Computational studies on protein-ligand docking

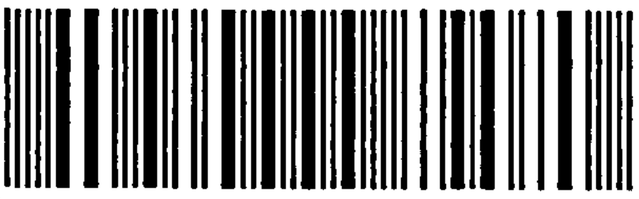
Thesis

Computational Studies on Protein-Ligand Docking

Thesis

M.Totrov • PhD

1999



Computational Studies on Protein–Ligand Docking

Thesis

M.Totrov • PhD

AUTHOR'S NO: M7283490
DATE OF SUBMISSION: 9 MARCH 1999
DATE OF AWARD: 10 MAY 1999

1999

EX12

RESEARCH DEGREES CENTRE

LIBRARY AUTHORISATION FORM

Please return this form to the The Research Degrees Centre with the two bound copies of your thesis to be deposited with the University Library.
All students should complete Part 1. Part 2 only applies to PhD students.

Student: Maxim Totrov PI: M7283490

Degree: PhD

Thesis title: "Computational Studies on Protein-Ligand Docking"

Part 1 Open University Library Authorisation [to be completed by all students]

I confirm that I am willing for my thesis to be made available to readers by the Open University Library, and that it may be photocopied, subject to the discretion of the Librarian.

Signed: Totrov Date: May 16, 1999

Part 2 British Library Authorisation [to be completed by PhD students only]

If you want a copy of your PhD thesis to be available on loan to the British Library Thesis Service as and when it is requested, you must sign a British Library Doctoral Thesis Agreement Form. Please return it to the Research Degrees Centre with this form. The British Library will publicise the details of your thesis and may request a copy on loan from the University Library. Information on the presentation of the thesis is given in the Agreement Form.

Please note the British Library have requested that theses should be printed on one side only to enable them to produce a clear microfilm. The Open University Library sends the fully bound copy of theses to the British Library.

The University has agreed that your participation in the British Library Thesis Service should be voluntary. Please tick either (a) or (b) to indicate your intentions.

[a] ☒ I am willing for the Open University to loan the British Library a copy of my thesis. A signed Agreement Form is attached.

[b] ☐ I do not wish the Open University to loan the British Library a copy of my thesis.

Signed: Totrov Date: May 16, 1999

Abstract

This thesis describes the development and refinement of a number of techniques for molecular docking and ligand database screening, as well as the application of these techniques to predict the structures of several protein-ligand complexes and to discover novel ligands of an important receptor protein.

Global energy optimisation by Monte-Carlo minimisation in internal co-ordinates was used to predict bound conformations of eight protein-ligand complexes. Experimental X-ray crystallography structures became available after the predictions were made. Comparison with the X-ray structures showed that the docking procedure placed 30 to 70% of the ligand molecule correctly within 1.5Å from the native structure.

The discrimination potential for identification of high-affinity ligands was derived and optimised using a large set of available protein-ligand complex structures. A fast boundary-element solvation electrostatic calculation algorithm was implemented to evaluate the solvation component of the discrimination potential. An accelerated docking procedure utilising pre-calculated grid potentials was developed and tested. For 23 receptors and 63 ligands extracted from X-ray structures, the docking and discrimination protocol was capable of correct identification of the majority of native receptor-ligand couples. 51 complexes with known structures were predicted. 35 predictions were within 3Å from the native structure, giving correct overall positioning of the ligand, and 26 were within 2Å, reproducing a detailed picture of the receptor-ligand interaction.

Docking and ligand discrimination potential evaluation was applied to screen the database of more than 150000 commercially available compounds for binding to the fibroblast growth factor receptor tyrosine kinase, the protein implicated in several pathological cell growth aberrations. As expected, a number of compounds selected by the screening protocol turned out to be known inhibitors of the tyrosine kinases. 49 putative novel ligands identified by the screening protocol

were experimentally tested and five compounds have shown inhibition of phosphorylation activity of the kinase. These compounds can be used as leads for further drug development.

Table of Contents

1. Introduction

Formation of non-covalent complexes is an essential part of almost any biological process. The remarkable complexity of the biochemical machinery of the living organisms would have been impossible without the ability of the participating molecules to recognise each other among thousands of other compounds simultaneously present in any cell. Specific binding between molecules is crucial in catalysis, signal transduction, molecular transport mechanisms, and determines the pharmacological effect of many drugs.

Better knowledge of the nature of molecular recognition on the microscopic level is important for our understanding of the normal and pathological processes in the cell and may help in such practical applications as drug design. X-ray crystallography has revealed detailed atomic descriptions of many individual proteins, nucleic acids and small biological molecules, as well as a number of structures of complexes. The Protein Data Bank (PDB) [1], where solved protein 3D structures are deposited, is growing by about 1000 new structures a year. Available structures of complexes can be analysed to discover the basic interactions and principles of molecular recognition, while the individual structures can be used in the prediction of unknown or novel complexes. First attempts to predict molecular interactions and design novel ligand utilised hand-made physical models of receptor sites and ligands [2]. Since manipulation of systems containing hundreds or thousands of atoms is necessary to simulate the binding process, the progress in numerical and computational approaches was essential for the advancement of macromolecular association studies. Computer simulations of molecular recognition were first attempted more than twenty years ago [3]. Considerable progress has been achieved in recent years, but reliability and precision of the existing complex prediction methods is still far from ideal.

1.1 Docking

1.1.1 Molecular docking problem

Prediction of the structure of a complex starting from the structures of individual molecules is commonly called the molecular docking problem. Structures of the protein-ligand and especially protein-protein complexes often show remarkable shape complementarity on the interface, suggesting the idea that the docking algorithms should search for such matching surfaces. Early approaches such as the original DOCK algorithm [3] used exclusively this geometric criterion. Both components of the complex were assumed rigid and the docking procedure searched for favourable mutual orientation using "sphere matching" [4], least-squares fitting of the surface patterns [5,6], Fourier-transform [7], distance-matrix matching [8] or "geometric hashing" [9]. Purely geometric approaches demonstrated certain success in recombining the structures of protein-protein complexes when the components were taken from the native complexed structure, which is a somewhat artificial starting point. In the more realistic cases, where the individual structures of the constituents were used, these techniques often failed to distinguish the correct orientation from the false positives [5]. High complementarity of the interacting surfaces in the native complexes is in part due to the "induced fit", e.g. the conformational change in the constituents of the complex upon binding, while individual structures often do not show the perfect matching expected in the complex. There are two general directions in which the simplistic geometric docking algorithms are being improved. First is the introduction of flexibility of ligand and/or receptor to reproduce or mimic the induced fit, and the second is the inclusion of binding determinants other than pure surface complementarity. Most attempts to introduce flexibility in protein-protein docking have so far been limited to "softening" of the geometric criteria which would allow a certain degree of penetration between the two interacting surfaces

[10,11]. Direct simulations with all-atom models may account for the flexibility more accurately and sometimes show promising results [12,13,14], but are often extremely computationally expensive. Whichever way the flexibility is introduced, it results in much greater ambiguity of the results of geometric docking, since many apparently good matches can be found. The multiplicity of solutions calls for additional criteria to select the correct answer. This lead to the inclusion in the docking protocol of the other binding determinants such as estimates of solvation free energy change or molecular mechanics energy [15], and ultimately, the approximations of the free energy change upon binding [16,17]. Most methods however still use simplistic measures during the generation of the bound conformations and than re-evaluate the putative solutions using more sophisticated potentials.

1.1.2 Docking as an energy optimization problem

Complexes considered in the docking studies are, in general, thermodynamically stable systems. Thus, the native bound conformation should represent the global minimum of the free energy of the system. Consequently, to find the docked conformation, the global minimum of the free energy function of the system has to be located. Since the precise evaluation of the free energy is difficult, one can try to use some approximation that would have a similar global minimum. From the energetic point of view, surface complementarity docking methods assume that the interaction energy is proportional to the contact area or other similar measure of the fit of two surfaces, possibly with some penalty for bad contacts (clashes). While this assumption may account reasonably well for van der Waals interactions and, to some extent, for solvation, it obviously disregards the energy contributions from specific pairwise atomic interactions such as hydrogen bond formation and electrostatics. Many recent docking studies try to incorporate these terms, often as the additional criteria to select the answer from many solutions generated by

geometric docking, either using force-field energy evaluation [15] or elaborate scoring functions [17,18]. In several works, physical energy terms were used throughout the algorithm [13,19].

Two major components are required for a successful prediction of the structure of the protein-ligand complex: an efficient global optimisation procedure which is capable of finding a global minimum for the strongly anisotropic function of dozens of variables, and a free energy approximation for the complex in solution which is computationally inexpensive to be used in the search procedure, yet sufficiently accurate to ensure the uniqueness of the native conformation. In the following two parts we will review the energy calculations and global optimisation techniques.

1.2 Energy terms

Energy calculations are at the centre of almost any molecular simulation technique. It is convenient and customary to divide the energy of the molecular system into a number of components, or *energy terms*. Below, five major components of the molecular interaction energy will be considered in greater detail.

1.2.1 Electrostatic Interactions

Electromagnetism is *the* fundamental force of biochemistry [20]. All processes on the molecular level can be described in terms of electromagnetic interaction combined with quantum mechanical and thermodynamic effects. While covalent and hydrogen bonding as well as van der Waals interaction all have an electrostatic nature, these interactions are complicated by quantum mechanics and it is often convenient to separate them from the longer-range electrostatic interactions. It is the latter type of interactions which is customarily referred to as

electrostatics in biomolecular structure. All proteins, and the large majority of ligands, contain polar atoms interacting strongly with each other and the solvent over a wide range of distances. For a charged amino-acid the strength of electrostatic forces may exceed by more than an order of magnitude the strength of van der Waals interaction [21].

The evaluation of electrostatic interactions in proteins was first attempted by Lingstrom-Lang in 1924 and Tanford and Kirkwood [22]. These macroscopic studies gave some qualitative insights, but only the availability of high-resolution protein structures and computer calculations allowed quantitative studies of protein electrostatics.

The largest problem in electrostatic calculations is the presence of highly polar solvent (water). In vacuum or in the uniform media the interaction between two charges can be simply described by Coulomb's law

$$E = k\frac{q_1 q_2}{\varepsilon R_{12}} \qquad (1.2.1.1)$$

where q_i are the charges, R_{12} is the distance between them, ε the dielectric constant and k is 332.0 when the charges are expressed in electron units, distance in angstroms and energy in kcal/mol. In an aqueous environment this relation has to be corrected to include the interaction of the charges under consideration with the large (virtually infinite) number of surrounding water molecules. Early attempts to simulate macromolecules without consideration of solvent screening ran into difficulties, such as DNA helices torn apart by electrostatic forces unless the electric charges were drastically reduced [23].

The straightforward and rigorous approach is to include explicitly a sufficiently thick layer of water molecules into the calculations. Obviously, it makes calculations heavier, but the principal difficulty of the explicit methods is that liquid water is an essentially dynamic environment. Any static placement of water molecules around the system under consideration would result in large errors, as the physically observed interaction with water is the result of averaging over a

large thermodynamic ensemble of the possible states of the solvent. Thus, to achieve accurate results one has to generate this ensemble by an extensive molecular dynamics simulation [54, 103]. While this might be the most rigorous approach to the solvation electrostatic calculations, it is impractical in many cases. The solvent effectively screens the interaction of the charges of the solute. Generally the farther from each other and the more exposed to the solvent charges are, the more their interaction is attenuated. This observation suggested simple corrections to the Coulomb law such as distance-dependent dielectric constant and charge-scaling. While it is somewhat *ad hoc* and doesn't take into account the interaction of the individual charges with the solvent (self-energy), distance-dependent dielectric constant $\varepsilon=\varepsilon_0 R$ is widely used because of its simplicity [104, 105]. This expression actually accelerates calculations of the energy and forces because they become dependent only on R^2 instead of R, eliminating costly square root calculations. Charge scaling was shown to improve the simulation results for such systems as DNA. While these crude approaches can hardly be used for quantitative evaluation of the properties of a macromolecule in solution, they keep the extra calculations to a minimum.

Alternatively, the solvent can be considered as a continuous medium of high dielectric constant. This treatment of the solvent is more computationally tractable than the inclusion of explicit water molecules. The electric potential obeys the Poisson differential equation

$$-\nabla(\varepsilon(\mathbf{r})\nabla\phi(\mathbf{r}))=\rho(\mathbf{r}) \qquad (1.2.1.2)$$

where ε is the dielectric constant (permittivity), ϕ is the electric potential, and ρ is the charge density. In a uniform medium it is equivalent to the Coulomb law, but the solution becomes more complicated when the space is divided into the regions of different dielectric permittivity. Analytic results exist only for special cases such as a sphere or a plane. Certain methods, for example the electrostatic image technique, utilise these analytic solutions to obtain relatively simple approximations of electrostatic energy under an assumption that the protein has

near-spherical shape, [106,107,40]. The precision of this approximation is obviously limited. A much more rigorous approach is to solve the Poisson equation numerically. Several techniques based on this idea were developed and are widely used in protein energy calculations [24].

1.2.2 Hydrophobicity

Transfer to the aqueous solution of a number of organic groups results in a free energy loss related to the ordering of water molecules around such groups which is known as the hydrophobic effect. The concept of the hydrophobic interaction was introduced by Kauzmann in 1959 [25]. This effect is similar in nature to the macroscopic surface tension. The hydrophobic interaction is a major driving force in the formation of most ligand-receptor complexes. For some ligands such as steroids the interaction is almost exclusively hydrophobic, and many other ligands are amphiphilic with hydrophobic groups binding into hydrophobic pockets of the receptor. By fitting the transfer free energies of hydrocarbons against the solvent accessible surface, the hydrophobic contribution was shown [26] to be proportional to the solvent accessible surface with fairly good precision. However, the coefficient of this proportionality is a subject to some controversy since it differs sharply from the microscopically observed value of the surface tension constant. The microscopic surface tension value derived from the transfer energies of aliphatic compounds is close to 30 cal/Å^2 while macroscopic hydrocarbon-water surface tension constant is ~75 cal/Å^2. Some attempts were made to explain the discrepancy by taking into consideration the curvature dependence of the surface tension and the difference of the molar volume of solute and solvent [27]. It remains to be seen if the division of the water-solute interaction into solvation electrostatics and hydrophobic components is the most adequate approach. Methods based on this partitioning were shown to reproduce successfully experimental data on transfer free energies for a large set of compounds [28].

However, alternative approaches to water-solute interaction evaluation were also developed, particularly a number of atomic solvation parameters (ASP) based methods. ASP methods differentiate the atoms of the solute into a number of types, each with a particular value of solvation energy surface density, generalising the surface tension. The underlying assumption is that the water-solute interaction can be partitioned into atomic contributions, which are proportional to the solvent accessible surface areas of the atoms. Popularity of the ASP approach is in part due to the simplicity and computational efficiency, while the drawbacks are that neither proportionality of the solvation energy to the accessible surface nor the partitioning of the solvation energy into atomic contributions can be rigorously justified, and are both largely *ad hoc* assumptions. Nevertheless, good agreement with experimental data can be achieved [29], which might in part be explained by the large number of adjustable parameters in the ASP models. It is questionable that these methods can perform well on a set of compounds which is much larger then the set used for the parameter adjustment.

1.2.3 Van der Waals interactions

The most generic type of interatomic force exhibits itself as a very strong repulsion at short distances and turns into relatively weak and quickly decreasing attraction as the distance between two atoms grows. It is commonly described by the "6-12" potential:

$$E_{VW}(R_{ij}) = -\frac{A_{ij}}{R_{ij}^{6}} + \frac{B_{ij}}{R_{ij}^{12}} \qquad (1.2.3.1)$$

where R_{ij} is the distance between the two atoms *i* and *j*. Parameters A_{ij} and B_{ij} depend on the types of atoms and are usually calculated using combination rules from the parameters for the identical pairs of atoms, which are in turn evaluated from quantum-mechanical or experimental data. Usually these parameters are derived along with the other components of the atomic interaction energy to form

so-called molecular mechanics force-fields, such as CHARMM [30], AMBER [31], MMFF [32] and ECEPP [33]. In this work we generally utilise ECEPP/3 parameterisation since it is optimised for internal co-ordinate representation.

While the $1/R^6$ form of the attraction term has strict quantum-mechanical basis, rigorous description of the repulsion term is more complicated. Alternative forms of the repulsion term have been proposed [32]. Fortunately, the interactions in biomolecular systems occur mostly in the range of inter-atomic distances where the attractive term is prevalent, and seem to avoid the strong repulsion, alleviating the problem of finding an exact description for the repulsion term.

Still, extreme sensitivity of the Van der Waals term to the small conformational changes makes its inclusion in the calculation of ΔG_{bind} problematic. This led a number of authors to simply omit the Van der Waals contribution in the binding energy, as it seems to introduce more noise than signal into the answer. Such omission is partly justified by the cancellation of ligand-receptor interactions in the bound state and the ligand-solvent/receptor-solvent interactions in the unbound state. One can assume that the overall number of interatomic contacts in the system remains nearly constant upon binding, resulting in the conservation of the total Van der Waals interaction energy. However, this approach leaves out entirely the dependence of the interaction energy on the quality of the interface. It might be of lesser importance for the prediction of the binding energy of the known complexes where the interface exhibits remarkable complementarity in most cases [34]. In the case of database scanning for novel ligands, the quality of the interface varies significantly and cannot be ignored in the evaluation of the binding energy. A possible compromise is to modify the Van der Waals potential so that it becomes less sensitive to the small deviations in atomic co-ordinates. We implemented this approach as described later.

1.2.4 Hydrogen Bonds

Hydrogen bond interaction is a specific attraction between polar hydrogens and a number of heavy atoms, which have lone electron pairs. From the large number of complexes whose 3-D structures have been solved it is known that many ligands form extensive networks of hydrogen bonds with their receptors, especially in cases of high specificity and high affinity binding. Hydrogen bonds also play an important role in protein folding, where their formation between the turns of the α-helixes and between the β-strands stabilises these essential secondary structure elements. Unfortunately, there seems to be no agreement so far about the adequate functional form for the hydrogen bonding interaction term and even the energetic value of an average hydrogen bond. Since its origin lays in the same electrostatic and quantum interactions as the origin of Van der Waals and electrostatic terms, hydrogen bonding is often included in the force field as a modification to the Van der Waals potential for the specific atom pairs [35,32]. The modification may only involve change in the parameters (MMFF), or a different functional form (10-12 instead of the standard 6-12 Van der Waals potential in ECEPP). Some force fields simply ignore hydrogen bonding in the hope that the electrostatic term will provide a sufficiently favourable contribution when positive hydrogen atoms and negative hydrogen bond acceptors are brought together. However, the charge distribution around the acceptor atoms is highly anisotropic since the lone electron pairs occupy sp^x orbitals, resulting in strong anisotropy of the HB interaction. High directionality of the HB interaction can also be observed in the solved structures of the proteins and protein complexes [36]. This anisotropy is largely ignored by pair-wise, atom-centric potentials used by the majority of the force fields. This omission may not lead to large errors as long as only naturally occurring conformations are considered, since they often already have optimal or sub-optimal configuration of hydrogen bonds. However, in the course of a simulation, such as docking, it may result in erroneous formation of hydrogen bonds of physically impossible geometries. Several forms of hydrogen-bonding term with explicit angular dependence were proposed [37,38].

1.2.5 Conformational Entropy

Binding of the ligand to the receptor usually imposes strong constraints upon its conformational freedom. In most cases, the bound state locks the ligand in a single conformation. Also, the surface side-chains of the receptor which are in contact with the ligand may no longer access some of their rotameric states. There is also a loss in translational and rotational degrees of freedom, which does not depend on the participating molecules and can be seen as a constant so long as we only consider 1-to-1 stoicheometry complexes. Binding may result in considerable decrease in the entropy which has to be included in the binding energy evaluation. As an illustration, one can consider the burial of one CH_2 group in an aliphatic chain. The loss of three rotameric states of the chain results in an entropy loss that adds $RT\ln 3 \approx 0.66$ kcal/mole to the free energy of the system, while the decrease in hydrophobic term is around -0.88 kcal/mole [39].

Exact determination of the entropy change would require extensive molecular dynamics simulations. Currently such simulations are too expensive computationally to use them routinely for a large number of putative complexes. An alternative approach is to assume that each free torsion rotation gives approximately the same contribution to the entropy. Then one can further assume that all torsions get locked upon binding, and use the simple count of torsions in the ligand multiplied by the constant per torsion contribution (usually 0.6 kcal/Mol). More sophisticated schemes attempt to evaluate the degree of conformational restraint imposed on each free torsion angle, e.g. according to the fraction of the accessible surface of the corresponding atom groups buried upon binding [40]. Furthermore, for the amino-acid residue side-chains statistical distributions of the χ angles can be derived from the database of known X-ray protein structures. These distributions can be used to calculate more exactly the entropy loss associated with locking of a particular side-chain in a single rotameric

state [40]. An alternative approach is to run a Monte-Carlo or molecular dynamics simulation of the components of the complex and of the complex as a whole to sample extensively the conformational space of the complexed and uncomplexed molecules and evaluate directly the change in the number of available states. Though such simulations give the most precise estimates of the entropy loss, they are prohibitively expensive computationally to be used routinely, especially in the case of molecular dynamics.

1.3 Conformational search techniques

An efficient global optimisation procedure is a key component of the docking protocol. Many approaches treat both ligand and receptor as rigid bodies [3,5,41]. Such treatment allows for rapid location of the optimal mutual orientation of the two molecules by special techniques (DOCK), but has limited applicability since the majority of small ligands are flexible and structural rearrangements occur in a number of receptors. To some extent, the limitations of the rigid-body docking can be circumvented if several low-energy conformations of the ligand are generated and then docked. The best solution can than be picked as an answer [42]. However, the number of conformations which have to be docked independently to achieve an accurate solution may become very large even for relatively small compounds. Therefore, many techniques try to treat the flexibility of the ligand more directly. The flexible ligand can often be partitioned into rigid fragments. For each fragment, rigid docking can produce a number of favourable orientations. Fragments are then reassembled into the original chemical structure ("Hammerhead", [43]) . Alternatively, one fragment is assumed to be essential for binding and placed in the active site first, then others are attached incrementally [44].

Two features of the protein-ligand energy landscape complicate the problem of the energy optimisation: high dimensionality and multiplicity of local minima.

High dimensionality makes the exhaustive search of the conformational space very computationally expensive. Large number of local minima makes rational determination of the global search direction virtually impossible and limits the usability of the derivatives to a small vicinity of one local minimum.

In order to deal with these difficulties, we use the technique of Monte-Carlo minimisation in internal co-ordinates.

1.3.1 Monte-Carlo

The term Monte-Carlo has been introduced by Metropolis and Ulam [45], with an allusion to the essentially random nature of such simulations. Monte-Carlo minimisation consists of three repetitive steps:

1. Random Jump. One or several variables in the system are changed randomly.

2. Local Minimisation. The energy of randomised conformation is optimised using a conjugate gradient or quasi-Newton technique to achieve a new local minimum.

3. Evaluation. The new conformation is accepted or rejected according to the Metropolis criterion: If the energy of the new conformation E_{new} is lower than the energy of the old one E_{old} , the new conformation is always accepted and used in the next iteration. Otherwise, it is accepted with the probability of

$P_{acc}=\exp(-(E_{new}-E_{old})/kT)$, where k is Boltzman's constant and T is the effective temperature of the simulation.

The Monte Carlo methods can be subdivided into local step and non-local step procedures, the former tending to make a random step in the vicinity of a current local minimum and the latter trying to jump to different minimum (in general not even to the neighbouring one) at each step. Rather sophisticated local step methods have been developed [108,109,110]. They find the appropriate search directions (related to the covariance matrix) in an attempt to make a step along low-energy valleys. However, these methods have limitations in their global

sampling capacity because they rely upon local harmonic approximation of the energy surface that is valid only close to the original conformation. This feature makes them adequate for sampling of the local environment of a certain conformation rather than for the large-scale searches.

In the alternative approach with non-local random steps, the main question is how to make the step, so that both the fraction of accepted random moves (the so-called acceptance ratio) and the performance are sufficient. High-dimensionality of protein systems clearly calls for much more efficient sampling algorithms than the existing ones. It has been established that a full local minimisation of each random step greatly improves the efficiency of the procedure [46,47]. However, some components of the energy, such as the solvation electrostatic energy, may have no derivatives and/or may be too computationally expensive for local minimisation. The double-energy MC minimisation scheme [13] circumvents this obstacle by using two sets of energy terms, one for the local gradient minimisation stage and another one for the Metropolis criterion evaluation stage in the MC step. Such division can be justified if the extra terms included for the Metropolis criterion are relatively "slow", insensitive to small conformational changes.

1.3.2 Internal co-ordinates

One of the principal difficulties in biomolecular simulations is the size of the system which often contains thousands of atoms. As a consequence, the conformational space has a very high dimensionality, complicating the search for the global energy minimum. The use of internal co-ordinates substantially reduces the number of variables defining the conformation of the system. The Cartesian description requires 3 variables (x,y,z) per atom. The internal co-ordinates description uses bond lengths, planar angles and torsion angles instead. Since bond lengths and planar angles are practically rigid under normal conditions, one can consider them as constants and only allow torsion angle changes (rotations around

the bonds), reducing the dimensionality of conformational space at least threefold. In practice, even greater reduction is achieved because at every branching point several atoms essentially share the same torsion angle (Fig 1.1 (a)).

The formal geometrical description to allow efficient manipulations of the multi-molecular system in internal co-ordinates with arbitrary subsets of free and fixed variables was introduced [13]. The technique represents the system as a directed treelike graph imposed on all atoms as well as on some auxiliary virtual atoms (Fig 1.1 (b)). Each atom in this basic description has three geometric parameters determining its position with respect to the preceding part of the tree. The parameters are bond length b, bond angle ω and torsion φ or phase ϕ dihedral angles for the main branch and side branches, respectively. The sub-trees of different molecules join in the starting triplet of virtual atoms which are fixed at the origin of the co-ordinate system and allow for standard treatment of all real atoms including the root atoms of each molecular sub-tree. When several internal variables are fixed (considered constant) a group of atoms may form so-called rigid-body, where mutual positions of the atoms involved do not change upon any changes of the remaining free variables. The concept of rigid bodies provides an important additional advantage for the energy calculations, since all pair-wise energy contributions from the atoms within a rigid body are constant. Such contributions often can be excluded from the calculations when only the relative energy change is important, improving the computational performance.

Throughout this study, we have applied the internal co-ordinate ideology in the representation of molecular conformations during local and global energy minimisation using the molecular simulation program suite ICM (Internal Co-ordinate Mechanics, Molsoft LLC).

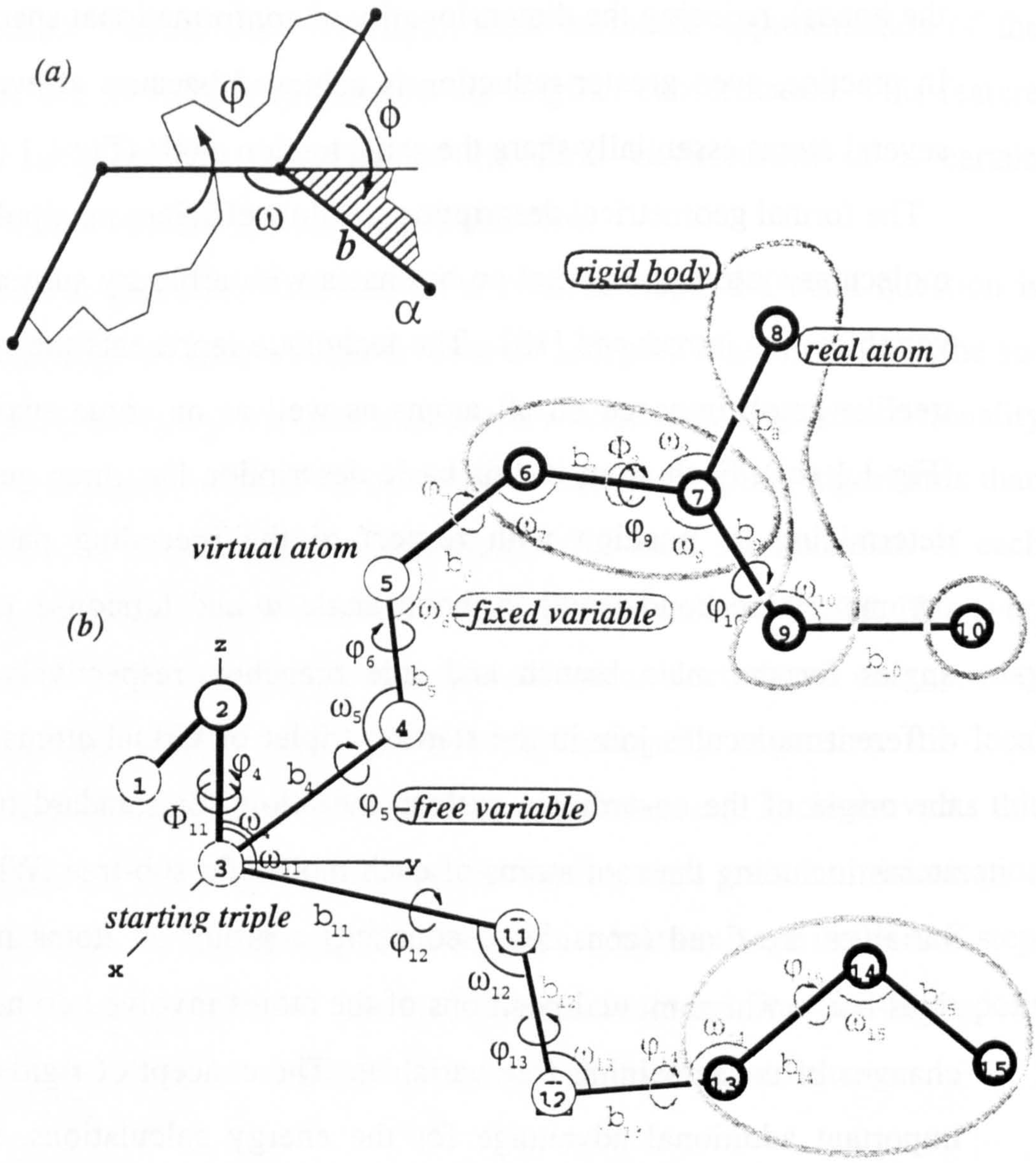


Fig. 1.1 (a) Four types of internal variables considered in ICM. (b) The ICM tree representing the geometry of multi-molecular arbitrarily fixed system and containing both real atoms and bonds (continuous lines) and virtual ones (dot-dashed lines). Atoms are numbered so that any atom in the directed graph starts a sub-tree with a continuous numbering. An arbitrary subset of free internal variables is shown in bold black characters, all the others being fixed (grey characters). The atomic regular directed graph is the basic one, the order of variables and rigid bodies following it. The numbering does not change as a result of re-fixation and redefinition of the rigid bodies. The attribution of the main (torsion) branch at the branching point is arbitrary and does not necessarily follow the atomic numeration.

1.3.3 Other approaches

Various global optimisation techniques were applied to the docking problem. Among the more popular is the genetic algorithm (GA), which was widely applied in protein folding simulations [111,112,113]. The idea of GA is to mimic the evolution process by manipulating "chromosomes", each containing a set of variable values defining a possible solution, e.g. a certain binding mode. The values inside the "chromosome" might be the rotatable torsion angles of the ligand and the variables defining the relative orientation of the ligand and receptor. The algorithm starts with a random "population" of chromosomes, from which new generations are produced by "mutations" and "crossovers", which involve, respectively, randomisation of some variables inside the chromosome or reshuffling of some variable values between two chromosomes. The best-fit "individuals" are preserved while others are discarded according to the fitness function. The assumption is that as the algorithm progresses, this strategy will find and keep the advantageous combinations of variable values, converging to the minimum of the fitness function. The GA docking was used fairly successfully to reconstitute a large number of known complexes [48], although no tests were undertaken to compare its performance with more conventional approaches such as MC.

Notably, Fourier-transforms were also used to locate the optimal geometric fit [7]. The method is efficient and attractively simple conceptually. Unfortunately it seems to be only applicable to a rather simplistic fitness function and can only optimise the three translational degrees of freedom. Rotations still need to be sampled by other means, i.e. systematic or random search. The Fourier-transform approach may be useful primarily in cases where the interacting molecules are very big, making other methods too expensive computationally.

Molecular dynamics (MD) simulation can be used as an optimisation method, and potentially it can provide a realistic picture of the binding process. However, MD is the most computationally expensive approach, and so far it is impossible to simulate the whole progress of the system from unbound components to the complex. The use of MD in docking is now limited to the simulations of the already bound complexes, where it is successfully used to predict various thermodynamic properties. Somewhat better performance can be achieved using so-called Brownian dynamics [49], which was applied to simulate long-range diffusion-like motions of the interacting macromolecules [50]. It is not clear if Brownian dynamics is capable of finding the final bound configuration of the complex.

1.4 Ligand discrimination

In the context of database screening, the goal of the docking simulation is two-fold: not only to predict the bound conformation for a particular ligand, but also to predict if any given compound would bind at all. The task of the discrimination procedure can be formulated as the assignment of a certain score to each compound in the database, reflecting the strength of its binding to the receptor. Ideally, such a score is the binding energy ΔG_{bind}.

1.4.1 Binding energy prediction.

As described in Chapter 1.2, a number of terms contribute to the free energy of the molecular system in solution and, subsequently, to ΔG_{bind}. Simple use of the molecular force-field energy has been shown to produce unsatisfactory results [15]. Van der Waals energy is extremely noisy because of the rigidity of the 6-12 potential, and its gain upon complexation is believed to be largely offset by the loss of the Van der Waals contacts with the solvent (water). These considerations suggest that it is advantageous to exclude Van der Waals energy entirely from the

evaluation of ΔG_{bind} [51,52]. Some approaches only include various forms of solvation energy [29] or solvation and electrostatic terms [51] plus a constant term to account for the translational and rotational entropy loss. Unfortunately, the precision of the calculations for these components of the energy is rather low, as the estimates for the values of their critical parameters, dielectric constant and surface tension, vary at least two-fold. The problem is further complicated by the fact that the individual terms are often much larger than the total ΔG_{bind}, negative contributions from hydrophobic and Coulombic terms being largely offset by the positive contribution from the desolvation of hydrophilic polar and charged atoms.

1.4.2 Discrimination score

Successful discrimination method should be able to rank the ligands in agreement with their experimental binding affinities, while remaining computationally manageable to be applied to a large number of potential complexes in an acceptable time. Most accurate techniques for binding energy prediction, such as free energy perturbation (FEP), achieve good agreement with the experiment but are rather slow [53,54]. Instead of attempting to predict the actual binding energy, one can try to generate a score which is a sufficiently good correlate of ΔG_{bind} to differentiate high-affinity ligands from non-binding compounds. Such a score can be based on the statistics of interfacial atomic contacts. Two studies applied the statistical approach with some success to the HIV protease inhibitors [114,115]. Earlier, a somewhat similar technique was proposed, splitting the binding energy into specific contributions from chemical groups such as COO^-, OH,CO etc., which were determined from existing binding energy data [55]. Other methods use matching of certain properties on the interface, such as hydrophobicity and hydrogen bond donors and acceptors (e.g. [56]).

2. Methods

2.1 Optimisation

2.1.1 Monte Carlo minimisation and conformational stacks

Throughout this study, MC minimisation was used as the main tool for the conformational search and optimisation. Main elements of the MC minimisation procedure in the internal co-ordinates are outlined above (sec. 1.3.1 and 1.3.2). To monitor the MC procedure and to introduce certain improvements into the basic MC protocol, the so-called conformational stack was maintained during the simulations [47]. The stack is a data structure containing the sets of variables defining certain conformations of the molecular system. The following algorithm was applied: At the start of the simulation, the stack is usually empty. After each accepted MC step, the generated conformation was compared to those stored in the stack. If the RMSD to all already accumulated conformations was above a certain cut-off value (*vicinity* parameter), the conformation was considered "new" and added to the stack unless the stack was full, i.e. the maximum number of conformations was already achieved. In that case the energy of the new conformation was compared to the energy of the worst (highest-energy) conformation in the stack. If the new conformation had lower energy, it replaced the old one, otherwise no modification of the stack was done. If the generated conformation was within the *vicinity* of one of the stored stack conformations, their energies were compared, and if the old one had higher energy it was replaced, otherwise no modification of the stack was done. For each conformation in the stack, the *number of visits* value was maintained. It was initialised as zero, increased each time the MC procedure generated a conformation within the *vicinity* from the particular stack conformation and reset to zero when it was replaced. Thus, at the end of simulation the contents of the stack could be

examined to discover the low-energy regions of the conformational space covered by the MC procedure and to determine how long the procedure has spent searching various regions. The information accumulated in the stack was also used to improve the efficiency of the search through the introduction of certain modifications to the basic MC procedure. To prevent the search from getting "stuck" in a single low-energy region, a limit on the *number of visits* for a stack conformation was imposed. After the limit was reached, a randomisation of variables much more drastic than the regular MC step was applied to force the system out of the over-visited region.

2.2 Boundary element numerical solution of the Poisson equation

2.2.1 Electrostatic interactions in solution

It is well recognised that electrostatic interactions have profound effects on macromolecular structure, folding and binding. Simplistic pair-wise Coulomb energy used in a number of molecular force fields proved to be inadequate in many cases since it does not account for the solvent effects.

The most rigorous approach might be an inclusion of explicit solvent (water) molecules into the system. Such calculations require addition of thousands of new atoms even for a moderately sized macromolecule and a long run of molecular dynamics is necessary to achieve even rather superficial sampling of the phase space for the added water molecules. Such sampling is necessary because the solvent molecules are not static and their thermal motion is essential for the electrostatic properties of the solvent.

As an alternative to explicit solvent one can use the continuous dielectric model. Instead of the discreet solvent molecules, the system is surrounded by a contiguous media of high dielectric constant. In this case, the electrostatic energy can be calculated as an energy of a set of point charges in a low dielectric constant

media surrounded by a high dielectric constant media. The molecular surface is usually taken as a boundary between the two. To calculate electrostatic potentials and energy in such a system, two major approaches have been developed: the finite difference method [24], and the boundary element method [64,63], as well as some approximations which use additional assumptions to achieve better performance, e.g. MIMEL [40], where quasi-sphericity is assumed in order to use the electrostatic image technique.

The finite difference method is currently the most popular approach. The major disadvantage of the method is that it requires manipulations of very large three-dimensional arrays since the properties such as electrostatic potential, charges and dielectric constant have to be represented on a grid. To achieve adequate precision, sub-Angstrom grids are required. For a medium-sized system of 50Å diameter and the grid mesh size of 0.5Å, the calculations involve the manipulation of several arrays of 100x100x100 = one million values each, which is both slow and memory-consuming.

An alternative approach is based on the mathematical observation that the solution of Poisson's equation for the system, where the space is divided into two regions of different dielectric permittivity, can be represented as the solution for a uniform medium if certain additional electrical charge density is distributed over the boundary between the regions. Since the electric field in the uniform medium obeys Coulomb's law, once the charge density on the boundary is known, electrostatic potentials and energy can be easily calculated. However, to find the boundary charge distribution, an integral equation has to be solved. The efficiency of the method depends to a great extent on the implementation of this solution.

2.2.2 Molecular surface

Since the continuous dielectric model involves the division of space into low-dielectric constant region (interior of the protein) and high-dielectric constant

region (solvent), it is important to define the boundary between the two. There are two types of surface widely used in the studies of solute-solvent interactions: *solvent accessible surface* which is displaced outward from the Van der Waals surface by the radius of the solvent probe [57,58]; and *molecular surface,* which is a smooth envelope touching the Van der Waals surface of atoms as the solvent probe rolls over the molecule [59]. In the electrostatic calculations the second type of surface is commonly used, as it represents the limits of the space which can be potentially occupied by solvent molecules. Several numeric algorithms were proposed for generating the approximations of the molecular surface, e.g. marching cubes [60]. Connolly [61] was the first to introduce a computer algorithm to generate the precise analytical molecular surface. In this work we used an improved contour-build-up algorithm to calculate the analytical molecular surface [62].

Tessellation of the molecular surface is an important part of the procedure and has a crucial effect on its performance and precision. The number of surface elements has to be kept as small as possible to make the calculations fast and reduce the amount of memory needed. On the other hand, the shape of the surface has to be adequately represented in order to achieve good precision. These contradictory requirements are hard to satisfy. Usually the surface is divided into triangles used as the boundary elements. One of the popular approaches essentially projects a pre-triangulated sphere onto the molecular surface [63]. While producing a relatively small number of surface elements, it will only generate satisfactory surface representation for a quasi-spherical molecule and any clefts, which are quite common for enzymes, may get severely distorted. Another approach is to use the detailed triangulated molecular surface [64], which guarantees good precision but is extremely slow and requires huge memory to store the matrix. Here we try to avoid both the excessive number of boundary elements and the oversimplification of the surface. Instead of directly using triangles as boundary elements, we combine them into relatively few patches of arbitrary shape. The

limited number of boundary elements keeps the calculations fast while their complex shape to a large extent preserves the precision.

2.2.3 Theory of the Boundary Element method

The basic task in the continuum dielectric solvation electrostatics calculation is to find the electric potential or field produced by a system of charges q_i in the region of space with dielectric constant ε_{in} surrounded by a medium with dielectric constant ε_{out} . The idea behind the boundary element method is to find an appropriate charge distribution on the surface of the dielectric boundary which would reproduce the same electric field in a uniform medium with dielectric constant ε_{in} . Once such a distribution is found, one can calculate the potential at any point from the Coulomb law since the dielectric is now uniform.

The electric field at an arbitrary point on the boundary should obey two conditions which can be used to deduce an equation for the surface charge density σ. The first condition is the continuity of the normal component of the electric displacement vector at any point on the boundary. If **n** is the normal to the boundary, $\mathbf{D}_{in}$ is the displacement just inside the boundary and $\mathbf{D}_{out}$ is the displacement just outside the boundary, than

$$\mathbf{D}_{in}\cdot\mathbf{n} = \mathbf{D}_{out}\cdot\mathbf{n} \qquad (2.2.3.1)$$

The second condition is for the discontinuity of the normal component of the electric field:

$$(\mathbf{E}_{out}-\mathbf{E}_{in})\cdot\mathbf{n} = 4\pi\sigma \qquad (2.2.3.2)$$

where $\mathbf{E}_{out}$ and $\mathbf{E}_{in}$ are the electric field vectors just outside and just inside the boundary, respectively. Combining these two equations and taking into account that $\mathbf{D}=\varepsilon\mathbf{E}$ we obtain the following equation relating $\mathbf{E}_{out}$ and σ:

$$\sigma = \left(\frac{\varepsilon_{in}-\varepsilon_{out}}{4\pi\varepsilon_{in}}\right)\mathbf{E}_{out}\cdot\mathbf{n} \qquad (2.2.3.3)$$

On the other hand, the electric field **E** can be calculated with the help of Coulomb's law from the known electric charges q_i and the charge density distribution σ:

$$\mathbf{E} = \sum_i \frac{q_i(\mathbf{r}-\mathbf{r}_i)}{\varepsilon_{in}|\mathbf{r}-\mathbf{r}_i|^3} + \oiint \frac{(\mathbf{r}-\mathbf{r}_s)}{|\mathbf{r}-\mathbf{r}_s|^3}\sigma_s ds \qquad (2.2.3.4)$$

where **r** is the radius-vector of the point where the electric field is being calculated, $\mathbf{r}_i$ are the radius vectors of the charges q_i, $\mathbf{r}_s$ and σ_s are the radius-vector and the surface charge density of an infinitesimal element of the boundary ds and the integral is taken over the whole boundary. This expression should not however be directly used in Eq. 2.2.3.2, because the surface integral has a discontinuity at the surface point. It can be shown that the value of the integral at the surface point and at the point infinitely close to it but outside the surface differ by $2\pi\sigma\mathbf{n}$, and for $\mathbf{E}_{out}$ one can have:

$$\mathbf{E}_{out} = \sum_i \frac{q_i(\mathbf{r}-\mathbf{r}_i)}{\varepsilon_{in}|\mathbf{r}-\mathbf{r}_i|^3} + 2\pi\sigma\mathbf{n} + \oiint \frac{(\mathbf{r}-\mathbf{r}_s)}{|\mathbf{r}-\mathbf{r}_s|^3}\sigma_s ds \qquad (2.2.3.5)$$

where **r** is now the radius-vector of a point on the boundary.

Now we can substitute $\mathbf{E}_{out}$ in Eq. 2.2.3.2 and obtain an integral equation for the σ:

$$\sigma = \left(\frac{\varepsilon_{in}-\varepsilon_{out}}{4\pi\varepsilon_{in}^2}\right)\sum_i \frac{q_i(\mathbf{r}-\mathbf{r}_i)\cdot\mathbf{n}}{|\mathbf{r}-\mathbf{r}_i|^3} + \left(\frac{\varepsilon_{in}-\varepsilon_{out}}{2\varepsilon_{in}}\right)\sigma + \left(\frac{\varepsilon_{in}-\varepsilon_{out}}{4\pi\varepsilon_{in}}\right)\oiint \frac{\sigma_s(\mathbf{r}-\mathbf{r}_s)\cdot\mathbf{n}}{|\mathbf{r}-\mathbf{r}_s|^3} ds$$

(2.2.3.6)

or

$$\sigma - \left(\frac{\varepsilon_{in}-\varepsilon_{out}}{2\pi(\varepsilon_{in}+\varepsilon_{out})}\right)\oiint \frac{\sigma_s(\mathbf{r}-\mathbf{r}_s)\cdot\mathbf{n}}{|\mathbf{r}-\mathbf{r}_s|^3} ds = \left(\frac{\varepsilon_{in}-\varepsilon_{out}}{2\pi(\varepsilon_{in}+\varepsilon_{out})}\right)\sum_i \frac{q_i(\mathbf{r}-\mathbf{r}_i)\cdot\mathbf{n}}{|\mathbf{r}-\mathbf{r}_i|^3}$$

(2.2.3.7)

To solve this integral equation numerically, we can break the boundary into fragments, or elements, and approximate the continuous surface charge

distribution by a set of surface charge density values, one for each boundary element. The integral equation 2.2.3.7 then turns into a linear equation system:

$$\mathbf{R}\sigma = \mathbf{e}\,, \tag{2.2.3.8}$$

where σ is the vector of surface charge density, matrix **R** only depends on the boundary shape and is defined by

$$R_{jk} = \delta_{jk} - \left(\frac{\varepsilon_{in} - \varepsilon_{out}}{2\pi(\varepsilon_{in} + \varepsilon_{out})}\right) \oiint_{S_k} \frac{(\mathbf{r}_j - \mathbf{r}_s)\cdot\mathbf{n}_j}{|\mathbf{r}_j - \mathbf{r}_s|^3} ds \tag{2.2.3.9}$$

and **e** is the vector defined by

$$e_j = \left(\frac{\varepsilon_{in} - \varepsilon_{out}}{2\pi(\varepsilon_{in} + \varepsilon_{out})}\right) \sum_i \frac{q_i(\mathbf{r}_j - \mathbf{r}_i)\cdot\mathbf{n}_j}{|\mathbf{r}_j - \mathbf{r}_i|^3} \tag{2.2.3.10}$$

Indexes j and k refer to the boundary elements, and $\mathbf{r}_j$ is the radius-vector of the point chosen to represent the 'centre' of the boundary element S_j.

The approximation of the integral in Eq. 2.2.3.9 can be obtained by using $\mathbf{r}_k$ instead of $\mathbf{r}_s$:

$$R_{jk} = \delta_{jk} - \left(\frac{\varepsilon_{in} - \varepsilon_{out}}{2\pi(\varepsilon_{in} + \varepsilon_{out})}\right) \frac{(\mathbf{r}_j - \mathbf{r}_k)\cdot\mathbf{n}_j}{|\mathbf{r}_j - \mathbf{r}_k|^3} S_k \tag{2.2.3.11}$$

However, this approximation is obviously not valid for the diagonal elements of the matrix, since it is singular for j=k. The simplest solution is to discard the integral for the diagonal elements completely, which would imply that individual surface elements are considered as flat, disregarding their curvature. Rashin and Namboodiri achieved some improvement in precision using finer tessellation of the surface for the calculation of the diagonal elements. Purisma and Nilar [65] have shown that if the surface is composed of interlocking spheres, one can deduce the diagonal elements of the matrix from the off-diagonal ones with the help of a certain normalisation condition. However, the molecular surface normally has many non-spherical (torroidal) elements. That, and especially the

imperfections of the surface may lead to erroneous estimates of the diagonal elements.

2.2.4 Implementation

A major problem of the boundary element method is the necessity of solving the linear equation system with the matrix **R** of the size NxN, where N is the number of boundary elements.

Unsatisfactory performance in terms of speed remains an obstacle for the wider usage of the boundary element method for macromolecular electrostatics calculations. In its simplest form, the method is only practical for the relatively small systems where a few hundreds of boundary elements are sufficient. Unfortunately, as the number of surface elements grows, the size of the matrix and especially the speed of the matrix inversion can make the boundary element electrostatic calculations impractical. The size of the matrix is proportional to the squared number of the surface elements (N), which means that one can not have more than a few thousand surface elements, i.e. 4000 elements require a matrix of 64 Mbytes in size. Also, the time required for the matrix inversion is proportional to the cube of N.

The second problem can be circumvented if, instead of the matrix inversion, one uses iterative solution of the linear system. When done properly, the iterative process usually converges in only a few (<10) steps. Making the size of the surface elements bigger helps to decrease the number of them, but quickly deteriorates precision. The characteristic size of the bumps and pits on the protein surface is close to the radius of an atom, which is about an angstrom for the hydrogens which are the majority of the surface atoms. If the surface is triangulated and the triangles are used as the boundary elements, these triangles should have the sides of less than an angstrom in length to retain any atomic

details. Bigger triangles may lead to exceedingly large errors: some atoms become very close to the surface and may even get outside of it.

To keep the number of boundary elements low we define our boundary element as an association of all triangles of the surface belonging to one atom. This allows us to have a detailed representation of the molecular surface shape with dozens of triangles per surface atom, while the number of boundary elements which are involved in the linear equation 2.2.3.8 is much lower and equals the number of surface atoms. The underlying assumption is that the variations of the surface charge density across the atomic surface patch are relatively minor.

Molecular surface was generated by the contour build-up algorithm [62] with subsequent triangulation of the three basic elements (convex spherical patches, torroidal saddles and concave spherical triangles). The algorithm also assigned each triangle to one of the atoms of the molecule. The assignment was used to group the triangles into patches of the surface used as boundary elements. The matrix elements R_{jk} were then calculated by summation of all contributions from the triangular components of patches *j* and *k*:

$$R_{j,k} = \sum_{l_j} \sum_{m_k} r_{l_j m_k} \qquad (2.2.4.1)$$

where l_j and m_k are indices of the triangles of the respective patches. The elementary triangular contributions $r_{l,m}$ were calculated according to Eq. 2.2.3.11. As mentioned above, special consideration has to be given to the diagonal elements r_{ll}. The expression 2.2.3.9 has to be written more precisely as

$$r_{ll} = 1 - \frac{1}{S_l}\left(\frac{\varepsilon_{in} - \varepsilon_{out}}{2\pi(\varepsilon_{in} + \varepsilon_{out})}\right) \oiint_{S_l} \oiint_{S_l} \frac{(\mathbf{r}_1 - \mathbf{r}_2)\cdot \mathbf{n}_1}{|\mathbf{r}_1 - \mathbf{r}_2|^3} ds_1 ds_2 \qquad (2.2.4.2)$$

If the surface element S_l is flat, the normal vector $\mathbf{n}_l$ is a constant and the integral becomes anti-symmetric with respect to $\mathbf{r}_1$ and $\mathbf{r}_2$, and since both integration variables cover the same surface, the integral is zero. Thus, the value of the integral reflects the curvature of the surface element S_l. As the calculation of the integral for the arbitrary shape of the element is difficult, we use the following

approximation: If the triangle belongs to the torroidal segment of the molecular surface, the value of integral is assumed to be zero since the average curvature of the surface is close to zero; if the triangle belongs to the spherical patch of the surface, the integral is estimated as $\pm S_i r_i^2 \sqrt{\pi}$ where r_i is the radius of the sphere and the sign is positive for the concave and negative for the convex triangles.

To solve the linear equation system, the conjugate gradients method was used [66]. Iterations converged after 8-12 steps. For an average-size 130 residue protein lysozyme, the total runtime to calculate the electrostatic energy was 13 second on R10000 195 MHz SGI Indigo computer. Surface calculation took 3 seconds, matrix preparation 4.5 seconds, conjugate gradients linear equation solution 4 seconds and final energy calculations 1.5 seconds. The time distribution shows that the technique is currently well-balanced, with no prominent bottlenecks. It was applied throughout this work for the electrostatic solvation energy calculations.

2.3 Deviation measure to rank docking solutions

To rank different conformations of a ligand of N atoms ($i=1,N$) docked to the receptor with respect to the known correct solution ($i'=1,N$) one may use a RDE (relative displacement error) measure which is related to the CAD [67] measure, but is much easier to calculate:

$$RelativeDisplacementError = 100 \left(1. - \frac{L}{N}\left(\sum_{i=1,N} \frac{1}{L + D_{ii'}}\right)\right),$$

where L is the scale parameter, N is the number of ligand atoms and $D_{ii'}$ is the deviation of the model atom i from the corresponding atom i' in the reference structure. The scale parameter defines the accuracy scale. Values of L between 1.5Å and 3Å are reasonable, since at these distances specific interactions of ligand atoms with the receptor atoms are significantly reduced and possibly replaced by

different interactions. From the above formula one can deduce the following properties of the proposed measure: if all the deviations are 0., RDE is 0%; if deviations are equal to *L*, RDE is about 50%; the same result may be achieved if half of the ligand atoms are predicted correctly (or deviate by much less than *L*), while the other half deviates by much more than *L*.

3. Flexible protein-ligand docking in full-atom representation.

Eight protein-ligand complexes were simulated using global optimisation of a complex energy function including solvation, surface tension and side-chain entropy in the internal co-ordinate space of the flexible ligand and the receptor side-chains [13,40]. The procedure uses two types of efficient random moves, a pseudo Brownian positional move [13] and a Biased-Probability multi-torsion move [40], each accompanied by full local energy minimisation. The best docking solutions were further ranked according to the interaction energy which included intra-molecular deformation energies of both receptor and ligand, the interaction energy, surface tension, side-chain entropic contribution and an electrostatic term evaluated as a boundary element solution of the Poisson equation with the molecular surface as a dielectric boundary. The geometrical accuracy of the docking solutions ranged from 30% to 70% according to the relative displacement error measure at a 1.5A scale. Similar results were obtained when the explicit receptor atoms were replaced with a grid potential.

3.1 Introduction

Theoretical prediction of the association of flexible ligands with protein receptors requires efficient sampling of the conformational space of a flexible ligand, a sufficiently accurate energy function and an efficient way to account for the receptor flexibility (see recent reviews [71,77,87,88]. Flexible docking schemes can be based on incremental construction of the docked conformation from separately docked rigid pieces [43,68,69] or on a limited discrete set of ligand conformations [42,70]. A molecular dynamics simulation of the entire continuously flexible ligand can be used to sample the conformational space of relatively small compounds [71,72,12,73]. Monte Carlo methods allow one to increase the sampling efficiency by making larger conformational rearrangements

[74,75]. Typically, sampling is performed by making random changes of one angle by a random value [75,76]. Caflish et al. [76] used Monte Carlo combined with local energy minimisation after each random change of a ligand torsion (receptor assumed to be rigid), as suggested by Li & Scheraga [46] for peptide structure prediction.

The continuous flexible docking procedure in internal co-ordinate space of both the ligand and the side-chains of protein receptor was first introduced in 1994 [13] and applied to predict the association of two α-helical peptides. This method attempted to globally optimise a rather complex energy function simultaneously with ligand and receptor rearrangements (each followed by local energy minimisation) rather then refine a set of solutions generated with rigid ligand molecules and with a simpler energy function. Later, the side-chain entropy and the MIMEL approximation of the solvation energy were added to the globally optimised objective function [40] , these terms being evaluated after each local minimisation as outlined in a 'double energy' scheme [13]. The ICM docking procedure correctly docked lysozyme and its antibody in full atom representations with flexible side-chain association and reached a discrimination of 19 kcal/mole between the correct lowest energy conformation and the closest false solution [19]. Later, the association of β-lactamase and its inhibitor [77,78] were correctly predicted with a similar energy discrimination gap, this time under blind prediction conditions.

Here, we apply the ICM docking method to small flexible ligands which are globally energy optimised together with the active site side chains using the double energy scheme. Additionally, we use an accurate boundary element solution of the Poisson equation to evaluate the 30 best docking solutions for each compound.

3.2 Method

3.2.1 Monte-Carlo conformational search

The ICM method describes both the relative positions of two molecules and their conformations by a uniform set of internal variables. Any subset of internal variables can be subjected to local or global energy minimisation procedures. Docking of flexible ligands into a flexible receptor requires three groups of free variables: positional variables of the ligand, intramolecular variables of the ligand and the torsion angles of the active site side chains (Figure 3.1). Flexible loops can also be sampled simultaneously with the ligand (e.g. in antibodies conformation of the loops is crucial for binding, see ref. 79 on immunoglobulin loop simulation). All the other variables are fixed to accelerate energy evaluation and sampling. In this study, the global minimisation procedure involved a random change of the internal variables followed by local energy minimisation (up to 100 steps of conjugate gradient minimisation) and selection by the Metropolis criterion (the temperature factor was set to 600K). Pseudo-Brownian random moves changed the position of the ligand molecule as a whole with a certain amplitude (here we used 2Å), as well as randomly rotated it around its centre of gravity by an angle close to the translation amplitude over the radius of gyration [13]. Internal torsion angles of the ligand were randomly changed one at a time, with an amplitude of 180°. Coupled groups of receptor side-chain torsion angles were sampled with biased probability moves [40].

Once the set of free variables was defined, the ICM global energy optimisation was performed from multiple starting points. Multiple starts were used to ensure convergence of the procedure. The number of starting points depends on the size of a ligand and here we used six random starting points. The energy optimisation routine consisted of the following iterative steps [13]:

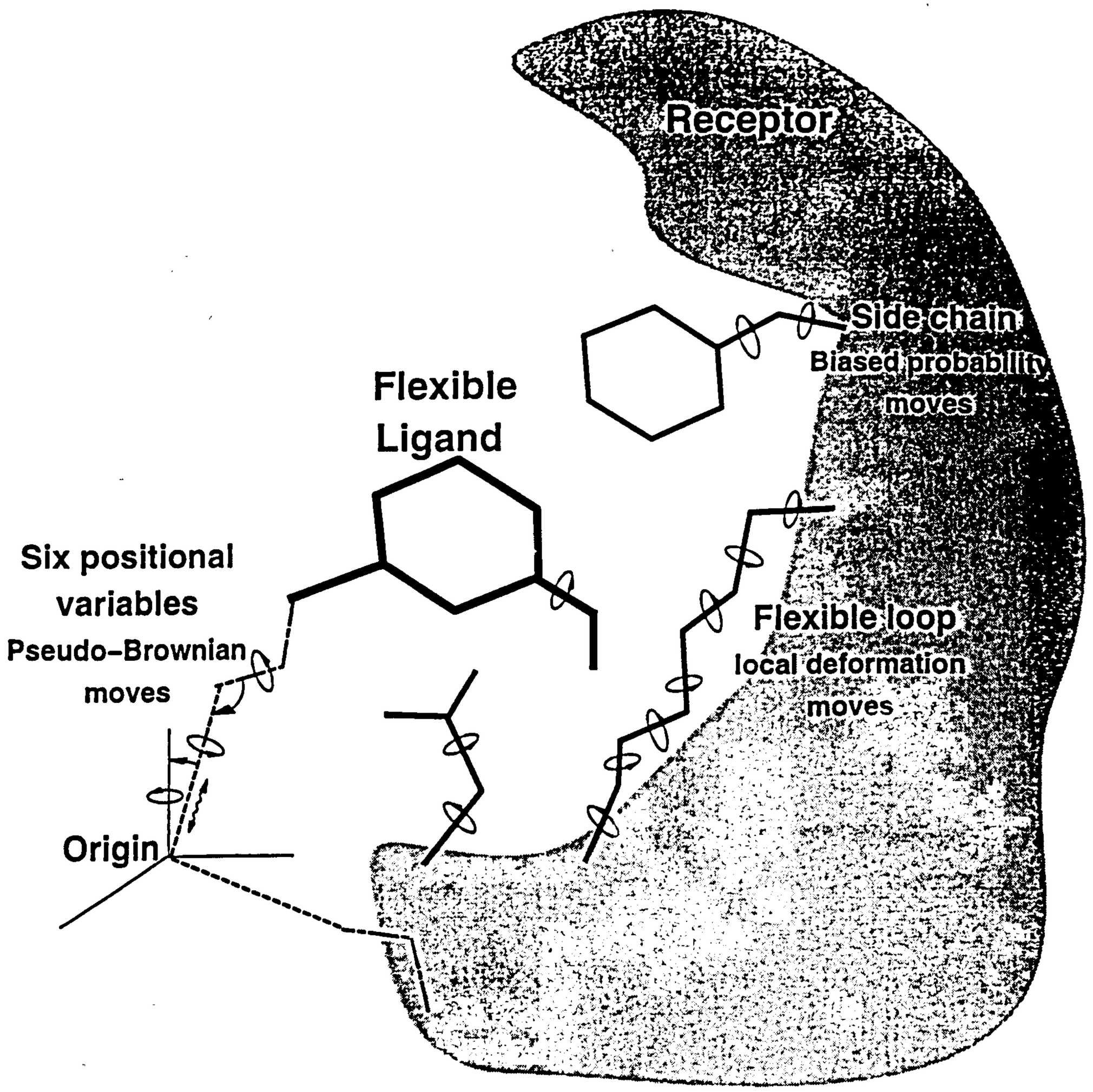


Fig 3.1. ICM docking set-up with flexible ligand and explicit flexible receptor. Most of the receptor variables are fixed, combining a large fraction of the receptor atoms into one rigid body.

- make a random conformational change of three possible types (Figure 1, loops were not considered here);
- perform local energy optimisation of the vacuum ECEPP3 energy [35] with a distance-dependent dielectric constant $\varepsilon=4r$;
- evaluate surface-based solvation energy and entropic contribution from the receptor side-chains and add it to the ECEPP3 energy;
- apply Metropolis et al. [45] selection criterion at a certain temperature T and make another step;

Geometrically different (as evaluated by the root mean square displacement of the ligand atoms with a threshold of 2.5Å) and low energy conformations were accumulated in the conformational stack [47].

3.2.2 Energy evaluation

During the MC runs, the energy was calculated using the ECEPP3 molecular force field, surface based solvation energy and entropic contribution. At the end of simulations, the conformational stacks were merged and the thirty best energy conformations were ranked with a more rigorous evaluation of the electrostatic free energy. Electrostatic free energy was calculated by a numerical solution to the Poisson equation using the boundary element algorithm [64] with $\varepsilon=4$ inside the molecules and $\varepsilon=80$ outside. Our implementation of the boundary element algorithm uses the accurate analytical molecular surface built by the fast contour build-up method [62] and is described in Chapter 2. The ECEPP charges [33] were used for the protein atoms. Since ECEPP doesn't provide the charge evaluation mechanism for an arbitrary chemical, charges of the ligand atoms were calculated with the quantum-mechanical program Gaussian [116].

3.2.3 Preparation of the individual initial structures

The techniques developed were tested on the docking prediction targets in the CASP-2 (Critical Assessment of Structure Prediction techniques) protein structure prediction contest. For the docking simulations, 8 ligand-protein complexes were proposed (Table 1). We made predictions for all eight complexes. For each of the targets, the co-ordinates for a complex of the protein with some other ligand(s) were found in the Protein structure Database (PDB), which allowed us to establish the approximate locations of the binding sites as a first step of the prediction. Next, three-dimensional models of the ligands had to be built. The chemical structures of the ligands were available in the form of connectivity tables. Since the experimental 3D co-ordinates for the ligands were not available, we built the models in the ICM program [13] from the fragments of the compounds found in the Cambridge Structural Database (CSD) [80] with known 3D structures. To find those, CSD was searched for the compounds with chemical structures similar to the chemical structure of the ligand. The third step was the assignment of partial charges to the individual atoms of the ligand, which were needed for the subsequent energy calculations. This was done with the help of the quantum-chemical program package Gaussian [116]. A CNDO hamiltonian was used to obtain the ligand atomic charges that are the most consistent with the standard ECEPP3 charges used for the protein molecule. The fourth and central step of the procedure was global energy optimisation of the ligand-protein complex. The ligand was placed in the vicinity of the binding site of the protein, and the system was subjected to the ICM docking procedure described above. During the procedure, torsion angles of the ligand and of the protein side-chains in a 7Å vicinity of the binding site were randomly changed. Each random change was followed by up to 100 steps of local conjugate-gradient minimisation. New conformations were accepted or rejected according to the Metropolis criterion using the temperature of 600K. Several independent Monte-Carlo runs of 300,000

energy evaluations were done for each ligand to ensure the convergence of the optimisation.

In the last step, putative solutions accumulated in the conformational stacks were re-evaluated using a more precise solvation electrostatic energy approximation based on the boundary element solution of the Poisson equation. The solution which scored best in this energy approximation was taken as the answer.

3.3 Results

3.3.1 Comparison of the predicted structures to X-ray results

For all 8 complexes, the best answers were submitted to the CASP-2 organisers. When the experimental structures became available, we were able to check the predictions. In most cases, the parts of the ligand inside the binding centre were predicted with good accuracy. Relatively large deviations occurred only for atoms outside the binding centre. We used RDE (relative displacement error) [67] as well as RMSD to evaluate our solutions. The results are summarised in Table 3.1. The high RMSD values for several complexes are somewhat misleading, because in fact only about half of the atoms of these ligands have large deviations, as the RDE measure correctly suggests. In the case of target 35, elastase/elastase inhibitor, the actual structure of the ligand has undergone chemical changes which were impossible to predict.

Table 3.I.

Results for the docking of eight ligands to their receptors evaluated by all heavy atom RMS deviation and the relative displacement error (RDE).

Target	Ligand	Receptor (PDB template code)	Site	Restraints	RMSD [B]	Fraction correct [C]
t13	methyl alpha-D-arabinofuranoside	Concanavalin A (5cna)	pocket	no	3.5	49.6%
t33	pentamidine	Pancreatic trypsin (2tbs)	pocket	tip [E]	9.27	51.7%
t34	amiloride,	Pancreatic trypsin (2tbs)	pocket	tip	4.2	48.1%
t35	SBA[A]	Pancreatic elastase (1inc)	covalent	chem. bond	10.6	31.2%
t36	SBB[A]	Pancreatic elastase (1inc)	covalent	chem. bond	10.7	35.6%
t39	Aica-Riboside Phosphate	Fructose bis-phosphotase (1fpd)	pocket	no	1.8	70.1%
t40	INH[A]	Pancreatic trypsin (2tbs)	pocket	tip	6.7	49.7%
t41	INI[A]	Pancreatic trypsin (2tbs)	pocket	tip	7.8	44.6%

[A] We use abbreviations suggested by the CASP2 organizers. SMILES strings of these compounds can be found at (http://PredictionCenter.llnl.gov/casp2/targets.html).
[B] Cartesian RMS deviation was calculated for all ligand heavy atoms with the receptor models superimposed.
[C] Fraction correct, or 100%-Relative Displacement Error, is calculated for all N heavy atoms of a ligand using this formula: $100\%\ (L/N)\ \Sigma(L+D_{ii})^{-1}$, where D_{ii} is the deviation of the model atom i from the corresponding atom in the reference structure, and the scale parameter $L=1.5$ L.

[D] Predictions were misled by the wrong chemical structure of the t35 ligand suggested for predictions.

[E] Tip indicates a distance restraint imposed on the carbon atom of the guanyl group.
Runtimes for simulations with fully flexible receptor side-chains and ligand varied from 5 to 15 hours.

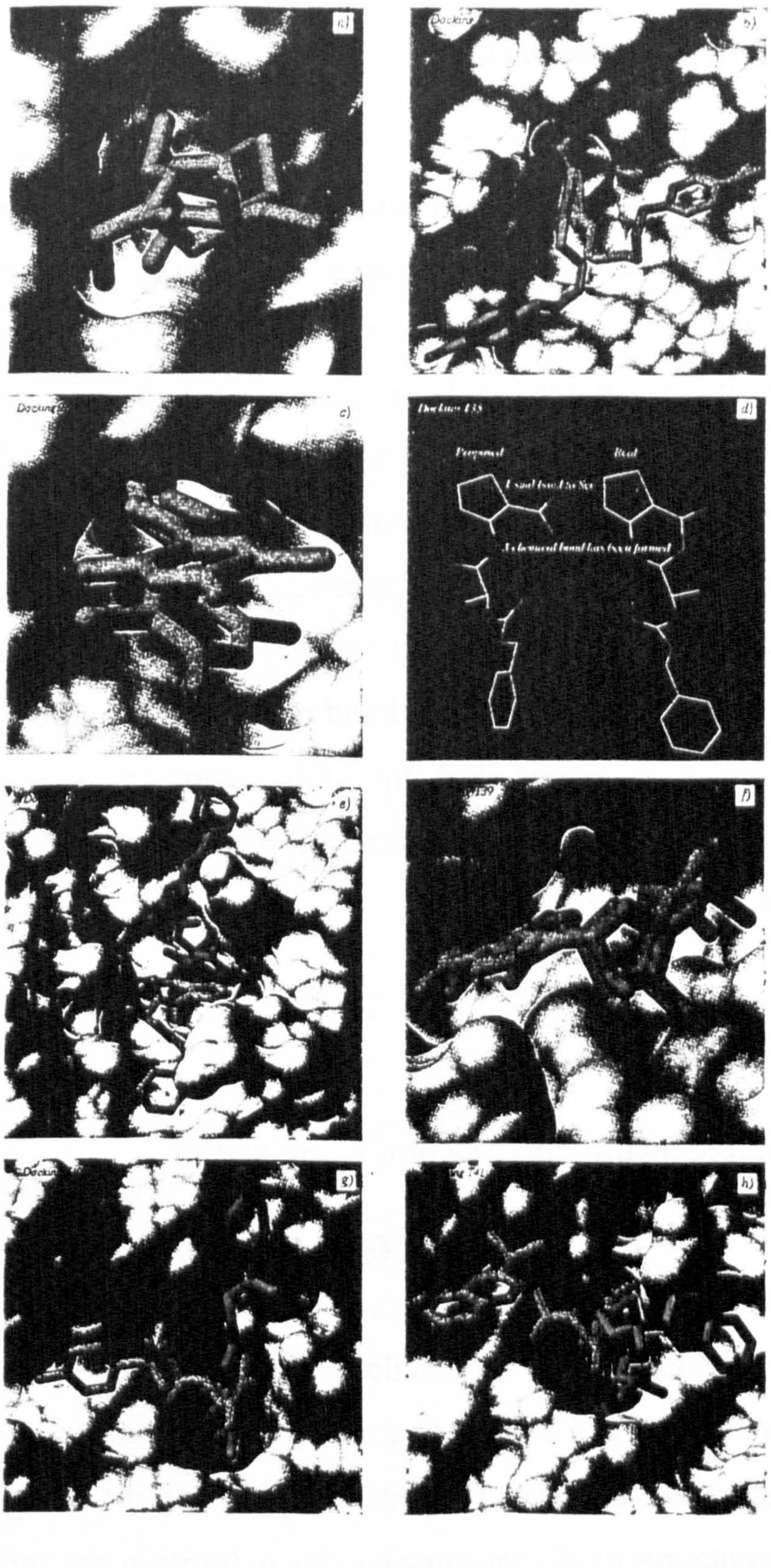


Fig. 3.2 Predicted docked conformations are shown in red and conformations determined by x-ray crystallography are shown in green. Analytical molecular surface of protein receptor was generated by the contour-build-up method [62].

3.3.2 Grid potential docking

Goodford [37] introduced the idea of pre-calculating the potentials produced by the receptor on a grid to accelerate radically the binding energy calculations. After the CASP-2 meeting, we attempted to predict the same complexes with the explicit receptor replaced by the grid potential. Four types of potentials were pre-calculated: the van der Waals potential for a hydrogen atom probe (1.0Å radius), the van der Waals potential for a heavy atom probe (generic carbon of 1.7Å radius was used) , an electrostatic potential from the receptor atoms, and the hydrogen bonding potential calculated as spherical Gaussians centred at the ideal putative donor and/or acceptor sites. Grid cell size was set to 0.5Å. Simulations took only about five minutes per compound and results similar to the results of the full-atom simulations were obtained. While this approach does not allow the explicit receptor flexibility, it might be preferred when the calculation speed is crucial, e.g. in database scanning.

3.4 Discussion

3.4.1 Flexible ligand and receptor optimisation

Accurate prediction of protein-ligand association requires inclusion of the ligand flexibility and protein surface flexibility in the docking procedure as well as precise evaluation of the interaction energy. The developed docking technique allows continuous and efficient sampling of internal torsion angles of the ligand and receptor side-chains as well as sampling of the variables which define the mutual orientation of the receptor and ligand within the same Monte-Carlo-based global optimisation framework. The pseudo-Brownian random moves differ from other schemes of random positional sampling, such as local minimisations from multiple starting points [81], or random translations and rotations (e.g. ref. [76]) and have the advantage of imitating local ligand rearrangements. The proposed

biased-probability sampling method for all surface side-chains in the vicinity of the active site is much more efficient than either discrete sampling [lea94] or changing one side-chain torsion angle at a time [76,81]. This method also can be used to sample the ligand if its conformational preferences in the form of continuous distributions are preliminary generated or evaluated using the database.

However, even if the global optimisation of the ligand/side-chain subsystem is fast and convergent, deformations of the backbone may still be crucial to docking with detailed atomic models. An adequate simulation of the backbone flexibility simultaneously with the ligand docking is still out of reach for the current computational approaches. To some extent, softening the potential (e.g. ref. [5,10]) or using an approximate grid potential [37,73,75], which is less steep than the realistic van der Waals repulsion, may be a practical way of overcoming this problem. Furthermore, simulations with the grid potential are much faster than the explicit flexible docking simulations and can be used for scanning large databases. Clearly, the choice between the explicit receptor model or the grid potential model depends on the docking problem and the available computer time and power.

In this work the receptor side-chains were sampled together with the ligand. Previously we found that for protein-protein docking this approach leads to a better discrimination between the correct and incorrect solutions [19,78]. It was unclear, however, that in this work the flexibility was essential.

3.4.2 Energy function

The energy function optimised by the procedure included a detailed vacuum energy complemented with the surface-based solvation and side-chain entropy. Since we intended to compare different conformations of the same ligand rather than binding affinities of different ligands, we did not estimate the ligand entropy loss [82]. However, inclusion of the side-chain entropy into global optimisation

[40] may be essential for discrimination between putative binding sites since these contributions can reach 2 kcal/mole/residue.

Numerical solution of the Poisson or the Poisson-Boltzmann equations (for review see ref. [20,83]) provides an accurate representation of the electrostatic solvation component of the ligand binding energy and can be added to the molecular mechanical force field to rank the docking solution [15,84,85]. We ranked the 30 best solutions using a more accurate evaluation of the electrostatic free energy calculated with the boundary element algorithm [63,64,86]. However, even these energies could not identify the correct positions of the solvent exposed parts of the long ligands. Technically, explicit water molecules could have been sampled together with the ligand, but explicit solvation can only be adequately considered within the framework of molecular dynamics.

3.4.3 Accuracy of predictions

Although the smaller compounds were predicted reasonably well, the relatively poor quality of prediction for the longer ligands suggests that the part of the ligand outside the binding pocket might not have a strong preference towards any one conformation. Presumably, the experimental structure in these cases is defined by a fine balance of energy terms which is still beyond the accuracy of the available energy approximations, or even perhaps by the crystallographic packing. The presence of many alternative configurations for such parts of the ligand molecule among the low-energy conformations accumulated during the simulations also suggests that the energy minimum for them is less well defined. Some of these alternative configurations are closer to the native conformation, but also have significantly higher energy then the lowest-energy conformation, suggesting that sampling of the conformational space of the ligand is sufficient. Further improvement in the free energy evaluation is necessary to achieve better docking precision for the weakly bound groups.

4. Ligand discrimination and fast flexible ligand docking using potential maps.

Discovery of new lead compounds is a crucial step in drug development. As the available computing power grows rapidly, virtual screening of the databases of chemicals becomes an increasingly viable alternative to direct experimental screening of hundreds of thousands of putative ligands. However, the ability of existing algorithms to distinguish high-affinity ligands from false positives remains low, with typical success rate of 1 active compound out of 10 to 50 selected by the screening protocol. We developed a novel approach to the derivation of binding potential by direct optimisation of its discriminative capability and derived an improved binding function using that approach. An exhaustive cross-docking of 23 receptors and 63 putative ligands extracted from high-resolution PDB structures of protein-ligand complexes was used as a benchmark. Ligands were diverse in size, from 12 to 84 atoms, and had a broad range of chemical properties and included sugars, fatty acids, phosphates, bases, heterocyclic and other compounds, which ensured the transferability of the scoring function to a larger variety of receptor/ligand pairs. Continuously flexible ligands were docked using the Monte-Carlo minimisation docking algorithm in the internal co-ordinate space. All complexes were subsequently evaluated and the best ligands for each receptor identified. The optimised scoring potential placed the native ligand first for 13 receptors and in all but two cases at least one native ligand was within the first 3 selected compounds.

4.1 Introduction

4.1.1.Overview

Automated identification of high-affinity ligands of various macromolecules (enzymes, receptors etc.) in a large database of compounds (virtual screening), may become a valuable tool in such applications as drug discovery. The basic idea of virtual screening is to find an optimal binding configuration for every compound in the database or in its large subset (docking, see revues ref. [71,77,87,88]), and then evaluate the bound structure by a discrimination procedure which should determine if one can expect high-affinity binding for the complex. A number of algorithms for docking and discrimination have been developed recently, such as DOCK [89], FLOG [42], and others (reviewed in ref. [90]). A number of successful applications of the scanning algorithms to discover the ligands for particular receptors were reported [90]. However, the success ratio is often relatively low. Therefore the improvement of the ligand discrimination remains an important issue. Attempts to achieve better discrimination were so far directed mostly towards finding a better approximation of binding energy ΔG_{bind} using the values of ΔG_{bind} for known complexes as a benchmark. In this work an alternative approach to the derivation of a sensitive discrimination algorithm will be described.

4.1.2 Ligand discrimination and its optimisation

The task of the discrimination procedure can be formulated as the assignment of certain score to each compound in the database, reflecting the strength of its binding to the receptor. Ideally that score would be the binding energy ΔG_{bind}.

Thus, every compound in the database has to be docked to the receptor and its ΔG_{bind} evaluated. Energetic terms contributing to the ΔG_{bind} are as follows:

(i) electrostatic interaction of the polar atoms with each other and the solvent,

(ii) hydrophobic term which accounts for the unfavourable interactions between certain groups and the solvent,

(iii) Van der Waals interactions,

(iv) formation of the hydrogen bonds,

(v) entropic contributions related to the loss of conformational freedom of the ligand and of surface side-chains of the receptor.

However each of the terms is also a source of errors, and for some of the contributions their magnitude is itself a matter of controversy, especially in the case of hydrophobic interaction, solvation electrostatics and hydrogen bonding interactions. Therefore, it appears legitimate to attempt an improvement of the discriminating potential by scaling the various contributions and subsequently optimising the weights. Moreover, since the goal in the derivation of an optimised potential is to achieve better discrimination, instead of trying to predict more accurately the actual binding energy of the known complexes, we will try to maximise the difference in the energy estimates between the binding and non-binding ligand-receptor pairs. Such an approach takes into consideration the numerous possible false pairs with low or no real affinity for each other which may have favourable apparent binding score, while traditional adjustment to the binding energy data only includes the (rather) high-affinity ligand-receptor pairs.

4.2 Materials and methods

4.2.1 Evaluation of discrimination potential performance

Optimisation requires a scoring function to evaluate the quality of a particular set of parameters. Our scoring function is based on a collection of known high-

resolution complexes from PDB from which a set of receptors and a set of ligands was extracted. For some receptors several complexes with different ligands were present in the data set, and vice-versa. Each ligand was then docked to each receptor, and the binding potential $E_{i,j}(\alpha)$ was calculated for the resulting complexes (α is the vector of current weights α_l for the potential terms, i and j are the indices of the receptor and ligand, respectively).

Parameter set evaluation function had the form

$$S(\alpha)=\sum_{i}\sum_{j_{bind(i)}}\max\left(E_{i,j}(\alpha)-\min_{k_{non-bind(i)}}(E_{i,k}(\alpha)),-10\right)$$

$$\alpha=(\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5) \qquad (4.2.1.1)$$

which is essentially the sum of the energy differences between the native ligands and first false positive for each receptor. Successful discrimination results in a negative contribution to $S(\alpha)$, while unsuccessful cases add positive penalty. The cut-off value of –10 kcal/mol was utilised to prevent the optimisation procedure from finding meaningless minima where one of the ligands would have exceptionally high separation at the expense of other ligand-receptor pairs.

4.2.2 Discrimination potential

Our binding potential consisted of the following terms:

$$E(\alpha) = \Delta E_{FF}+\Delta E_{EN}+\alpha_1 N_{at}+\alpha_2\Delta E_{HB}+\alpha_3\Delta E_{SE}+\alpha_4\Delta E_{EL}+\alpha_5\Delta E_{SO} \qquad (4.2.2.2)$$

ΔE_{FF} is the force-field energy which included inter- and intra-molecular Van der Waals interactions and torsion energy for the ligand calculated with ECEPP/3 parameters[10]. Since ECEPP/3 only has parameters for amino-acid atom types, the atoms of ligands were assigned closest chemically similar atom types. Because of its extreme rigidity, Van der Waals potential in its standard 6-12 form may

introduce large noise in the energy function. For inter-molecular interactions we therefore used a modified smoother form of the potential with most of the repulsive part truncated. Truncation was achieved by the following transformation of the original value of Van der Waals potential:

$$E_{VW} = \begin{cases} E^0_{VW}, if & E^0_{VW} \le 0 \\ \dfrac{E^0_{VW} E_{\max}}{E^0_{VW} + E_{\max}}, if & E^0_{VW} > 0 \end{cases} \qquad (4.2.2.1)$$

This expression ensures smooth transition from undistorted form of Van der Waals potential in the negative range of values to increasingly attenuated form in the positive range, asymptotically approaching E_{max} cut-off value. E_{max} was chosen on the basis of preliminary tests to be 1.5 kcal/mole. Lower values sometimes result in severely clashed docking solutions as the Van der Waals repulsion is no longer able to compete with attractive terms, primarily electrostatic. This and other potentials were pre-calculated on a grid to accelerate energy evaluation during the simulations. The grid cell size was set to 0.5Å.

ΔE_{EN} is the entropic contribution, which was estimated as 0.6 kcal/mol/K times the number of free torsions in the ligand. ΔE_{HB} is hydrogen bonding term which was calculated using Gaussian-type potential positioned around the centre of each lone electron pair of the hydrogen-bond acceptors:

$$E_{HB} = E^0_{HB} e^{\frac{(r - r_{hb})^2}{d^2_{HB}}} \qquad (4.2.2.2)$$

The peak interaction energy E^0_{HB} was assumed to be 2.5 kcal/mol as an average of various estimates, and the radius of the interaction sphere d_{HB} was assumed to be 1.4Å, allowing for about 30° to 40° deviation from the ideal geometry in accordance with observations in X-ray structures. r_{hb} is the radius-vector of the interaction centre, which was placed 1.7Å from the atom. In case of hydrogen atoms the centre was placed along the axis of the covalent bond attaching the hydrogen to the rest of the molecule. In case of heavy sp^2 atoms, one (for nitrogen)

or two (for oxygen) centres were placed at the angle of 120° to the existing covalent bond. For sp^3 oxygen and sulphur, two centres were placed in tetrahedral geometry, at 109° to the existing covalent bonds and to each other.

We used a combination of two types of electrostatic terms: distance dependent electrostatics with $\varepsilon=4r$ (ΔE_{EL}) and solvation electrostatics calculated by the boundary element solution of the Poisson equation with dielectric constant set to 4 for the inside of the molecule (ΔE_{SE}) [63,84,85,86].

The hydrophobic term ΔE_{SO} was calculated as proportional to the buried hydrophobic surface with the free energy density of 30 cal/mol/Å^2. To accelerate calculations, a grid-based form of the hydrophobic potential was developed. The fragments of the solvent-accessible surface were generated using the modified Shrake and Rupley algorithm [13,58]. The algorithm produces dots, which evenly cover the surface. The hydrophobic potential on the grid was then calculated as:

$$E_{SO} = E_{SO}^{0} e^{\frac{d_{surf}^2}{d_w^2}} \qquad (4.2.2.3)$$

d_{surf} is the distance to the closest point of the hydrophobic surface, and d_w is effective radius of the hydrophobic interaction which was set to the diameter of the water molecule 2.8Å. The value of E^0_{SO}=3 kcal/mole was chosen to reproduce the surface tension of 30 cal/mol/Å^2 for extended hydrophobic surfaces in test cases. The non-physical term proportional to the number of atoms in the molecule was introduced after the preliminary tests showed a bias of the energy function towards bigger ligands. This trend might be explained by our use of the "softened" Van der Waals term which can result in artificial extra Van der Waals attraction roughly proportional to the number of the atoms.

4.2.3 Docking

To generate docked conformations used in the binding potential evaluation, the Monte-Carlo minimisation technique in the internal co-ordinates was utilised. The protocol similar to the one used in Chapter 3. The energy function during the MC run included all the terms described above with the exception of solvation electrostatics since it is too computationally expensive to be used during the docking procedure. Single solvation electrostatics energy evaluation may take up to a minute of CPU time, depending on the receptor size, while the entire MC docking simulation takes from 1 to10 minutes, depending mostly on the ligand size. The adaptive length of the MC runs was used, with the limit on the total number of steps proportional to the size (number of atoms) of the ligand: $N_{MCsteps}=50*N_{LigAtom}$. Similarly, an adaptive length of local minimisations during the MC run was used: $N_{LocMinSteps}=25+N_{LigAtom}$. The factors in these relations were established empirically from the convergence and efficiency considerations.

4.2.4 Optimisation

The set of 23 receptors (Table 1.) and 63 putative ligands (Table 2.) was extracted from high-resolution PDB structures. The structures were selected according to a number of criteria: All structures at resolutions worse than 2.0Å were discarded since large errors in the receptor co-ordinates could result in poor docking and recognition for reasons unrelated to our study. Several complexes had the ligand bound covalently to the receptor and were also discarded since the prediction of such chemical reactions is beyond the scope of our approach. We also omitted complexes where metal ions were directly involved in the protein-ligand interaction since the force field used in the simulations did not provide for adequate modelling of such atoms. Some of the ligands in the set were retained from the structures of the receptors not used in the simulations for one of the reasons described above. We nevertheless kept these ligands to enrich the set and evaluate the discrimination protocol under more stringent conditions. For a

Table 4.1 Receptor structures used in docking simulations and recognition experiments

PDB code	Receptor name
188l	Lysozyme mutant
1ake	Adenylate kinase
1ars	Aspartate aminotransferase
1erb	Retinol binding protein
1fkh	FK506 binding protein
1fnd	Ferredoxin reductase
1gar	Glycinamide ribonucleotide transformylase
1gca	Glucose/galactose-binding protein
1hmr	Fatty acid binding protein
1hsl	Histidine-binding protein
1icm	Intestinal fatty acid binding protein
1lst	Lysine-, arginine-, ornithine-binding protein
1mai	Phospholipase c δ-1
1mdq	Maltodextrin-binding protein
1mrg	α-momorcharin
1mrj	α-trichosanthin
1nsc	Neuraminidase
1rcf	Flavodoxin
1sre	Streptavidin
2dri	D-ribose-binding protein
2tbs	Trypsin
4dfr	Dihydrofolate reductase
Fgf[1]	Tyrosine kinase of FGF receptor

1.The X-ray coordinates were kindly provided by S. Hubbard.

number of receptors, structures of several complexes with different ligands were available. In all such cases we used a single receptor structure in recognition experiments with all ligands. Hydrogen atoms were added to all X-ray structures using the hydrogen placement algorithm of ICM software [13]. Electric charges were assigned to the atoms of the ligands using bond-charge increment algorithm from MMFF94 force field [32]. Each ligand in the set was docked to every receptor using the flexible Monte-Carlo docking procedure with potential maps as implemented in ICM software [13,40,91]. For every complex, all energy terms

were evaluated and the resulting three-dimensional ($N_{ligand} \times N_{receptor} \times N_{term}$) set of data was utilised in the subsequent optimisation.

To find the optimal set of parameters, the zero order simplex minimisation algorithm ("amoeba") was implemented [66]. To ensure that the global minimum was found, 10 independent optimisation runs from random starting points with 10 re-starts of the "amoeba" algorithm were conducted. All of the optimisation runs went quite far from the our initial guess of parameter vector α=(0. 1. 1. 0. 1.) (see eq. 1) which included most terms with the weight of one and omitted the size factor and distance-dependent electrostatics.

Table 4.2 Ligands used in docking simulations and recognition experiments

Code	Source (PDB code)	Name
ela	1hmr	elaidic acid
ola	1hms	oleic acid
ste	1hmt	stearic acid
mya	1icm	Myristate
mtx	4dfr	Methotrexate
ddf	1dyj	5,10-dideazatetrahydrofolate
dzf	1dyh	5-deazafolate
fol	1dyi	Folate
ffo	1jom	folinic acid
ben	2tbs	Benzamidine
amc	1tng	Aminomethylcyclohexane
fba	1tnh	4-fluorobenzylamine
pbn	1tni	4-phenylbutylamine
pea	1tnj	2-phenylethylamine
pra	1tnk	3-phenylpropylamine
tpa	1tnl	Tranylcypromine
oxe	188l	o-xylene
ind	185l	Indole
i4b	184l	Isobutylbenzene
pxy	187l	p-xylene
n4b	186l	n-butylbenzene
bnz	181l	Benzene
den	183l	Indene
bzf	182l	Benzofuran
gtt	1hnl	glutathione
pgh	1tpb	phosphoglycolohydroxamate
etr	1erb	n-ethyl retinamide
fen	1fel	fenretinide
rea	1fem	Retinoic acid
aze	1fen	axerophthene
hab	1sre	haba
mhb	1srg	3'-methyl-haba
dmb	1sri	3',5'-dimethyl-haba
nab	1srj	naphthyl-haba
icl	1inc	benzoxazinone
gis	8est	guanidinium isocoumarin
ibr	9est	guanidinium isocoumarin
baa	1elg	n-(tert- butoxycarbonyl-alanyl-alanyl)-o-(p-nitrobenzoyl) hydroxylamine
t44	1eta	thyroxine

u89	1gar	burroughs-wellcome inhibitor 1476u89
fad	1fnd	adenosine-2',5'-diphosphate
ap5	1ake	inhibitor ap5a
fmn	1rcf	flavin mononucleotide
adn	1mrg	adenine
mal	1mdq	maltose
gal	1gca	galactose
rip	2dri	beta-d-ribose
nag	1nsc	N-acetyl-D-glucoseamine
st1	1ivd	4-(acetylamino)-3-hydroxy-5-nitrobenzoic acid
lys	1lst	Lysine
his	1hsl	Histidine
e6c	1ppp	Inhibitor e64-c
e64	1aec	Inhibitor e64
clm	3cla	chloramphenicol
plp	1ars	pyridoxal-5'-phosphate
pmb	1xzc	para-sulfurousphenyl mercury
sbx	1fkh	(1r)-1-cyclohexyl-3-phenyl-1-propyl (2s)-1-(3,3-dimethyl- 1,2-dioxopentyl)-2-piperidinecarboxylate
i3p	1mai	inositol trisphosphate
Mil	1tlm	milrinone
Mpd	1nco	2-methyl-2,4-pentanediol
Sia	1nsc	sialic acid
Dan	1nsd	2,3-dehydro-2-deoxy-n-acetyl neuraminic acid

4.3 Results

4.3.1 Grid docking

51 complexes with known structures were predicted. 35 predictions were within 3Å from the native structure, producing correct overall positioning of the ligand, and 26 were within 2Å, giving fairly detailed picture of the receptor ligand interaction. Individual cases are further analysed:

Lysozyme mutant complexes o-xylene, indole, isobutylbenzene, p-xylene, n-butylbenzene, benzene, indene and benzofuran. Eight small aromatic and heteroaromatic compounds were docked into the cavity left in the core of the protein after the mutation of bulky residues for smaller ones. This is an example of low-specificity primarily hydrophobic binding. All but one compound were docked in the conformation closely resembling native, with RMSD under 1Å. In the case of isobutylbenzene, the benzene and isobutyl groups are exchanged, resulting in rather large RMSD of 4.5Å (Fig. 4.1).

Adenylate kinase complex with the inhibitor ap5a. This is one of the biggest ligands in the set. Despite the size of the problem, the docking shows remarkable precision, with both adenosine moieties nicely docked into their pockets and somewhat higher deviations in the less specific central poly-phosphate region. Overall RMSD is 0.98Å (Fig. 4.2).

Aspartate aminotransferase complexed with pyridoxal-5'-phosphate. Due to the Shiff base formation, in the native structure one of the carbon atoms of the ligand is very close to the lysine 258 nitrogen atom. This type of interaction is not permitted by the docking technique used and resulted in considerable deviation of the docked structure. However, the overall binding mode is well conserved in the model generated, with RMSD 2.37Å (Fig. 4.3).

Retinol binding protein complexes with n-ethyl retinamide, fenretinide, retinoic acid and axerophthene. The correct binding mode was found in all cases, though quality of the prediction varied, with RMSD values from 0.92Å to 2.21Å (Fig. 4.4).

FK506 binding protein complex with (1r)-1-cyclohexyl-3-phenyl-1-propyl (2s)-1-(3,3-dimethyl- 1,2-dioxopentyl)-2-piperidinecarboxylate. The solution found by the docking procedure was correct on the large scale, while the positioning of several groups was rather approximate, RMSD 2.27Å (Fig 4.5).

Ferredoxin reductase complex with FAD. The experimental complex structure actually contains adenosine-2',5'-diphosphate as well as FAD. The docking

procedure correctly identified the flavin binding pocket, but failed to place the adenosine moiety which doesn't seem to interact strongly with the receptor (Fig. 4.6)

Glycinamide ribonucleotide transformylase complex with Burroughs-Wellcome inhibitor 1476u89. While there are minor deviations throughout the predicted structure of the ligand, all the functional groups are placed correctly. RMSD 2.19Å (Fig. 4.7).

Glucose/galactose-binding protein complex with galactose. The predicted structure has correct orientation with a minor overall shift, mostly due to the inappropriate choice of the isomer of histidine 152 in the receptor. RMSD 1.27Å (Fig. 4.8).

Fatty acid binding protein complexes with elaidic, oleic and stearic acids. For these complexes the prediction was one of the worst. They seemingly lack strongly localised specific interactions, and the ligands are extremely flexible molecules. The flexibility makes proper sampling of their conformational space difficult. While parts of the long aliphatic chain do follow the same course in the predicted structures, other parts deviate strongly. The RMSD ranges from 3.49Å to 6.69Å (Fig. 4.9).

Histidine-binding protein complex with histidine. While overall position of the ligand is predicted correctly, the imidazole ring is flipped. This flaw stems again from the problem of the isomeric states of histidine. Better docking could have been achieved if two alternative states were tried. RMSD 1.68Å (Fig. 4.10).

Intestinal fatty acid binding protein complexes with myristate and oleate. In contrast to the other set of fatty acid complexes, the predictions here rather closely resemble the native structures, even though the crystallographic structure of the oleate complex has considerable disorder, with three alternative positions for the carboxyl group. RMSD's are 1.79Å and 1.46Å (Fig. 4.11).

Lysine-, arginine-, ornithine-binding protein complex with lysine. Remarkably good prediction, with RMSD of 0.61Å (Fig. 4.12).

Phospholipase c δ-1 complex with inositol trisphosphate. Rather poor prediction, possibly due to the complicated hydrogen-bond network formation in the native complex. RMSD 5.03Å (Fig 4.13).

Maltodextrin-binding protein complex with maltose. Though many details of the hydrogen-bonding network are not reproduced in the model, the two sugar rings are placed accurately. RMSD 0.92Å (Fig 4.14).

α-momorcharin complex with adenine. Incorrect docking, possibly due to insufficiently good treatment of hydrogen bonds. Interestingly, the five nitrogens of the adenine molecule are fairly well superimposed, but with incorrect order. RMSD 3.51Å (Fig. 4.15).

α-trichosanthin complex with adenine. Very good prediction, RMSD 0.42Å (Fig. 4.16).

Neuraminidase complexes with sialic acid, 2,3-dehydro-2-deoxy-n-acetyl neuraminic acid and 4-(acetylamino)-3-hydroxy-5-nitrobenzoic acid. All three predicted structures are fairly good, with only minor deviations. RMSD's are 0.93Å, 0.75Å and 0.83Å respectively (Fig 4.17).

Flavodoxin complex with flavin mononucleotide. Flavin moiety docking precision is excellent, while the sugar and phosphate have higher deviation, possibly due to the relatively loose contact of these groups with receptor. RMSD 1.2Å (Fig. 4.18).

Streptavidin complexes with 2-((4'-hydroxyphenil)-azo)benzoate (HABA), 3'-methyl-HABA, 3',5'-dimethyl-HABA and naphthyl-HABA. Two of the complexes (the first and the last) are predicted correctly, while in the other two the phenyl rings are reversed. RMSD's are 1.50Å, 7.24Å, 7.53Å and 0.76Å respectively (Fig. 4.19).

D-ribose-binding protein complex with beta-d-ribose. Good prediction with RMSD 0.56Å (Fig. 4.20).

Trypsin complexes with inhibitors benzamidine aminomethylcyclohexane 4-fluorobenzylamine 4-phenylbutylamine 2-phenylethylamine 3-phenylpropylamine

tranylcypromine. While for the first three inhibitors prediction was satisfactory, for the last four precision was rather poor. In these complexes, the amino group penetrates much deeper into the active site than for the benzamidine-like compounds. It was the structure of the receptor from the trypsin-benzamidine complex that was used to generate the predictions. Possibly, the induced fit steered the prediction in all cases towards benzamidine-like binding. RMSDs are 1.91Å, 1.27Å, 1.94Å, 3.05Å, 2.79Å, 2.60Å and 2.11Å respectively (Fig. 4.21).

Dihydrofolate reductase complexes with methotrexate 5,10-dideazatetrahydrofolate 5-deazafolate folate folinic acid. The overall position of the ligand is correct in all complexes, but placement of the individual chemical groups vary from fairly good in case of metatrexate to unsatisfactory for the folinic acid. RMSDs are 1.81Å, 2.84Å, 2.48Å, 3.45Å and 5.00Å respectively (Fig. 4.22).

Tyrosine kinase of FGF receptor complex with Sugen inhibitor. Good prediction with RMSD 0.76Å (Fig. 4.23).

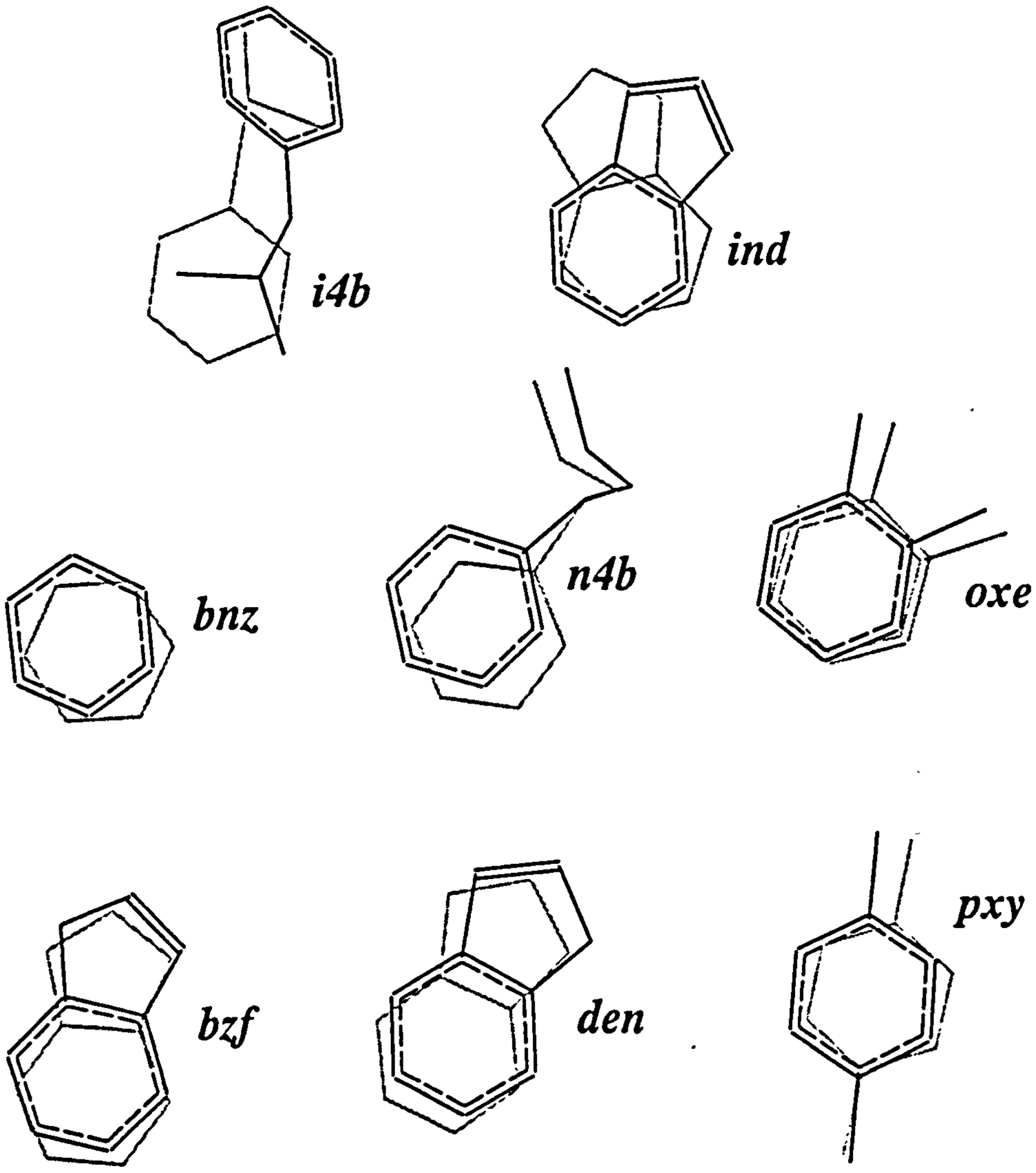


Fig. 4.1 Predictions and experimental conformations for the lysozyme mutant complexes o-xylene, indole, isobutylbenzene, p-xylene, n-butylbenzene, benzene, indene and benzofuran

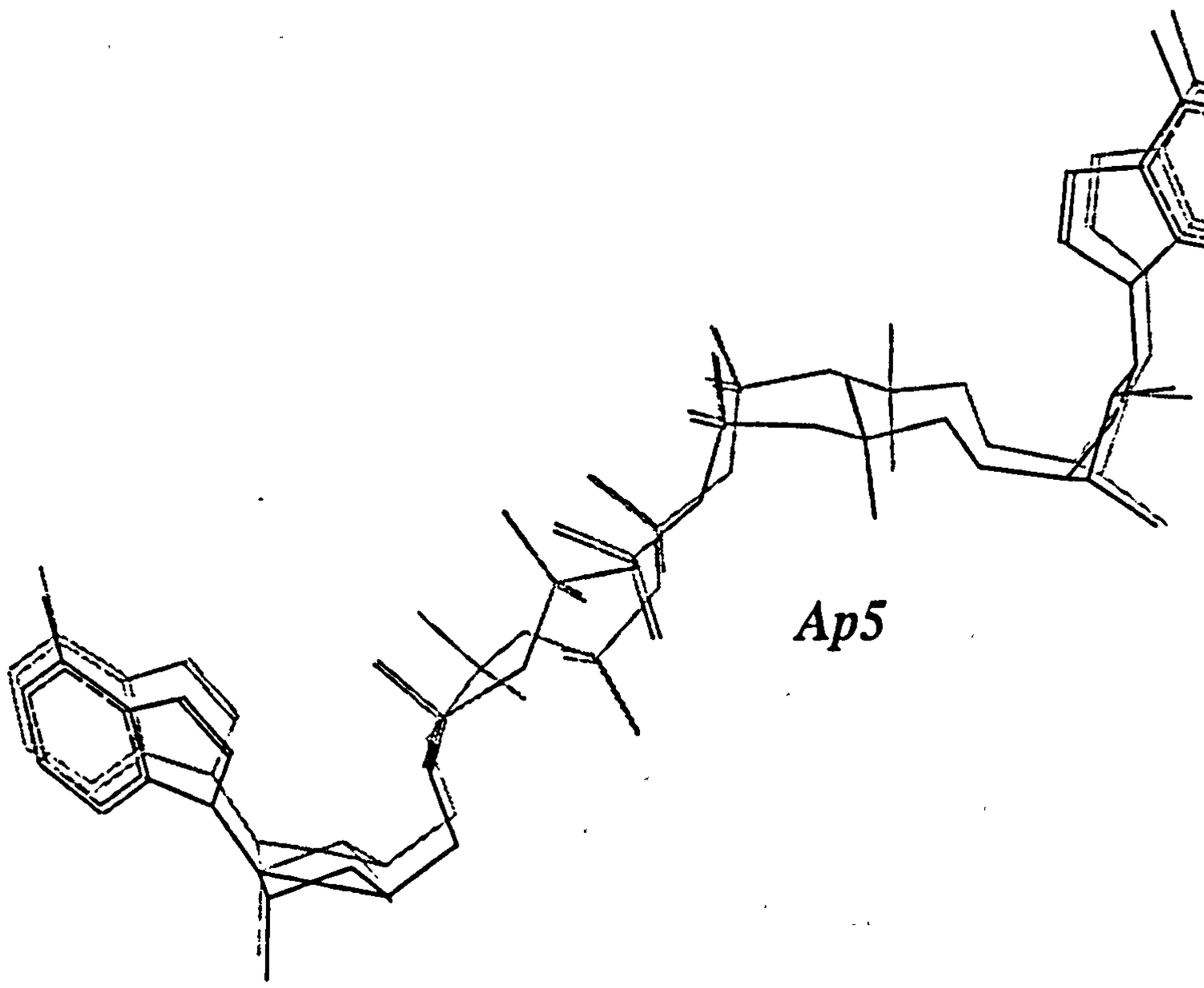


Fig. 4.2 Predictions and experimental conformations for the adenylate kinase complex with with the inhibitor ap5a.

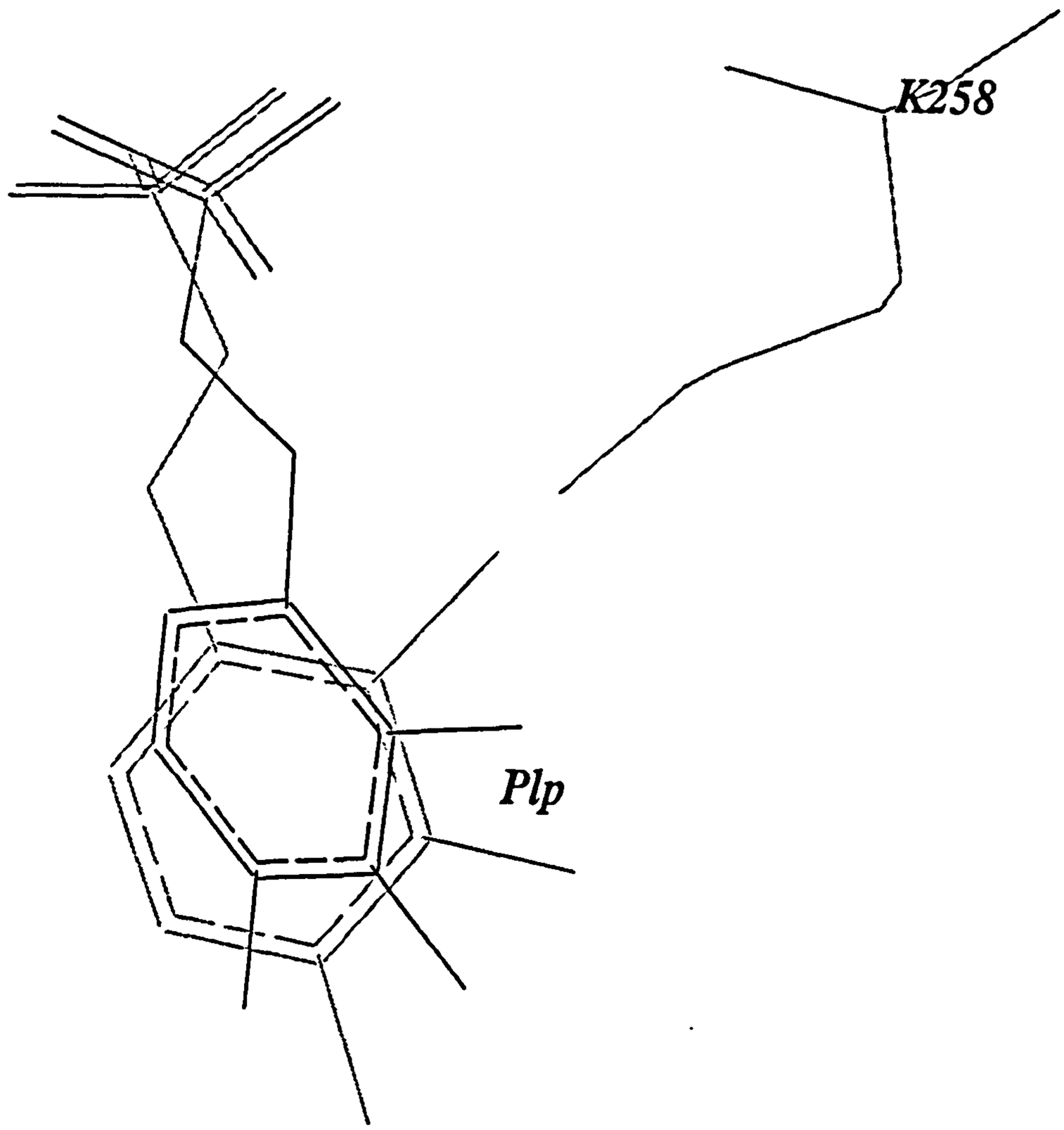
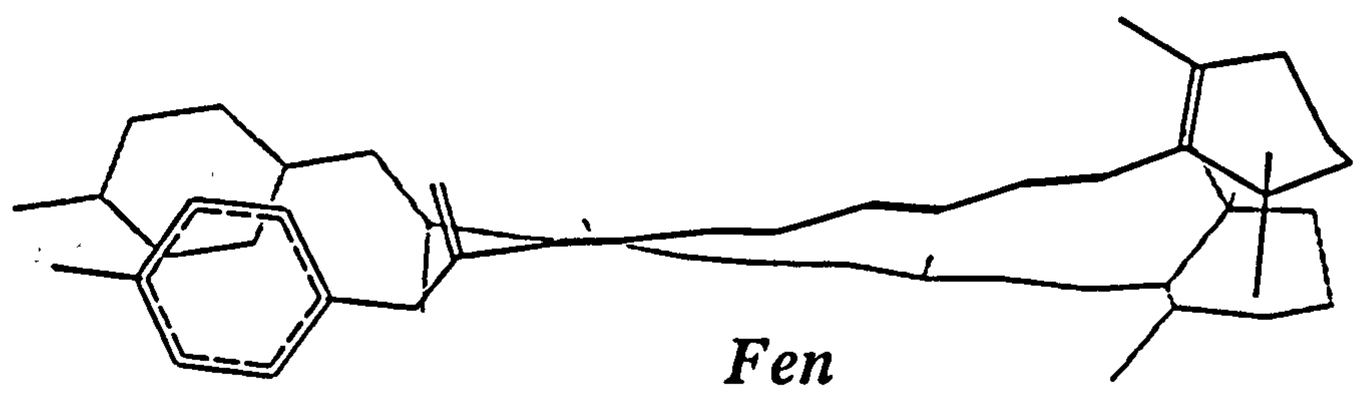
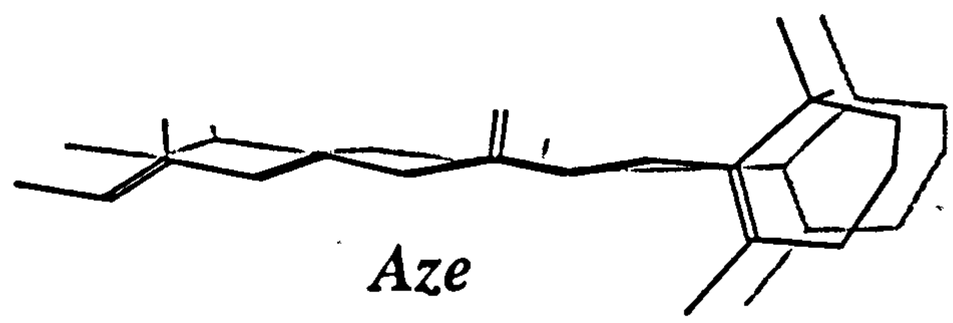


Fig. 4.3 Predictions and experimental conformations for the aspartate aminotransferase complexed with pyridoxal-5'-phosphate.





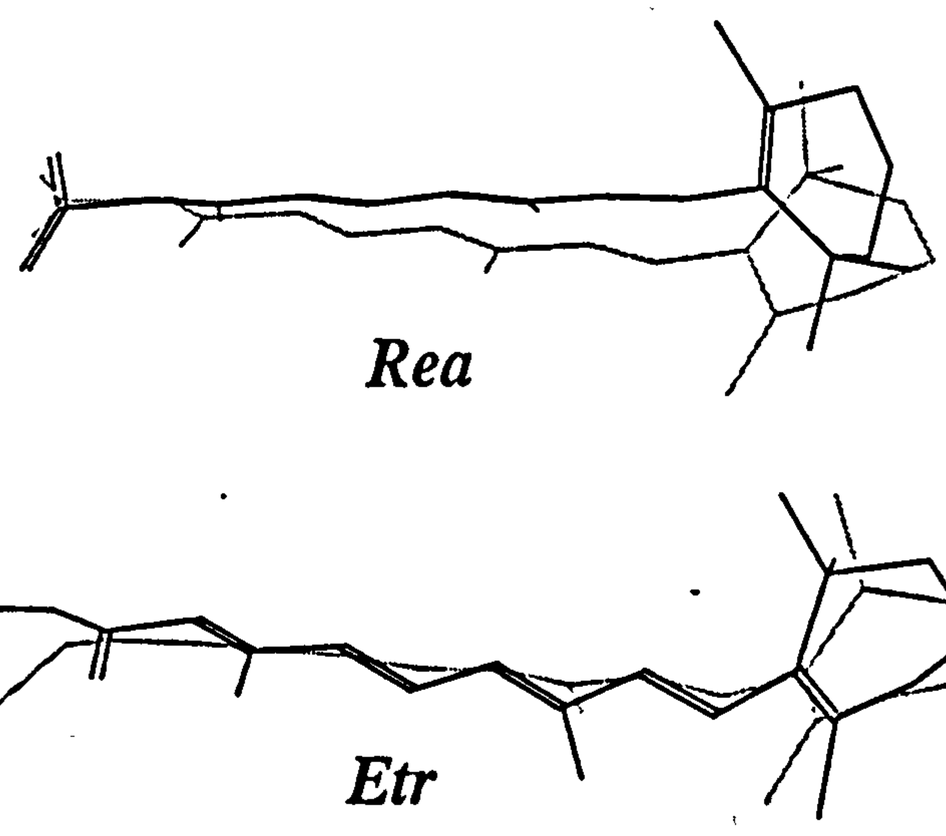
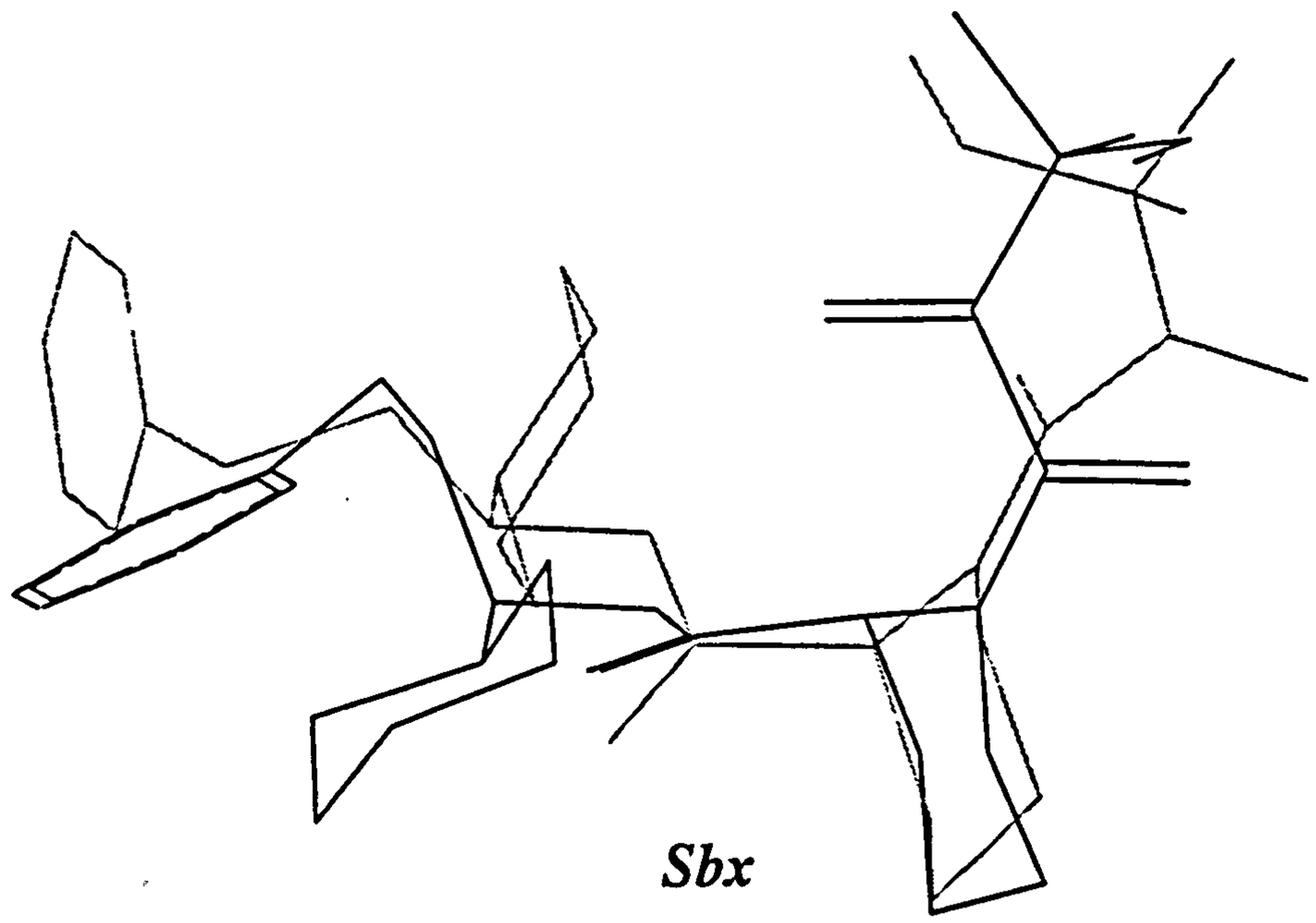


Fig. 4.4 Predictions and experimental conformations for the retinol binding protein complexes with n-ethyl retinamide, fenretinide, retinoic acid and axerophthene.



Sbx

Fig. 4.5 Prediction and experimental conformation for the FK506 binding protein complex with (1r)-1-cyclohexyl-3-phenyl-1-propyl (2s)-1-(3,3-dimethyl- 1,2-dioxopentyl)-2-piperidinecarboxylate.

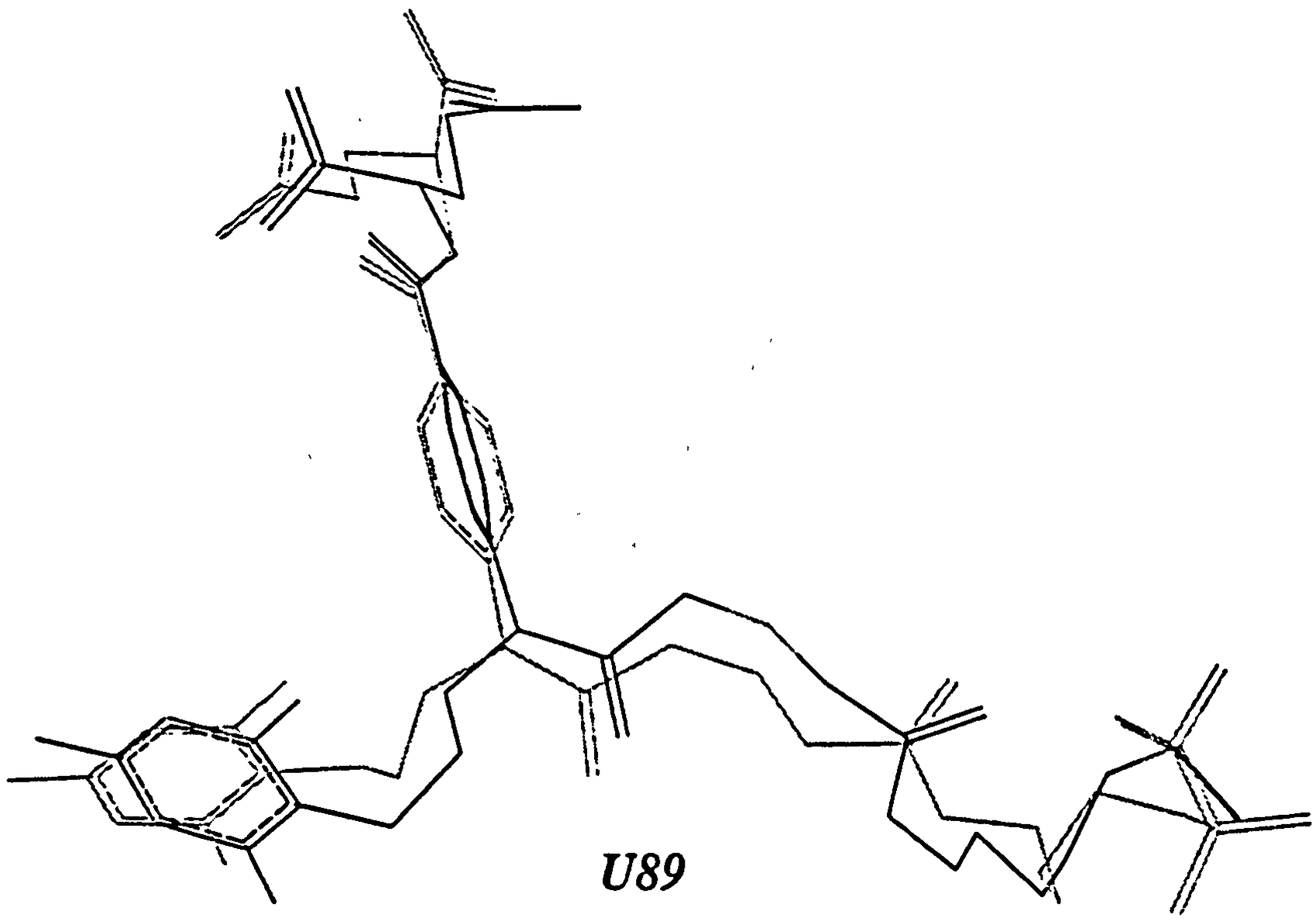
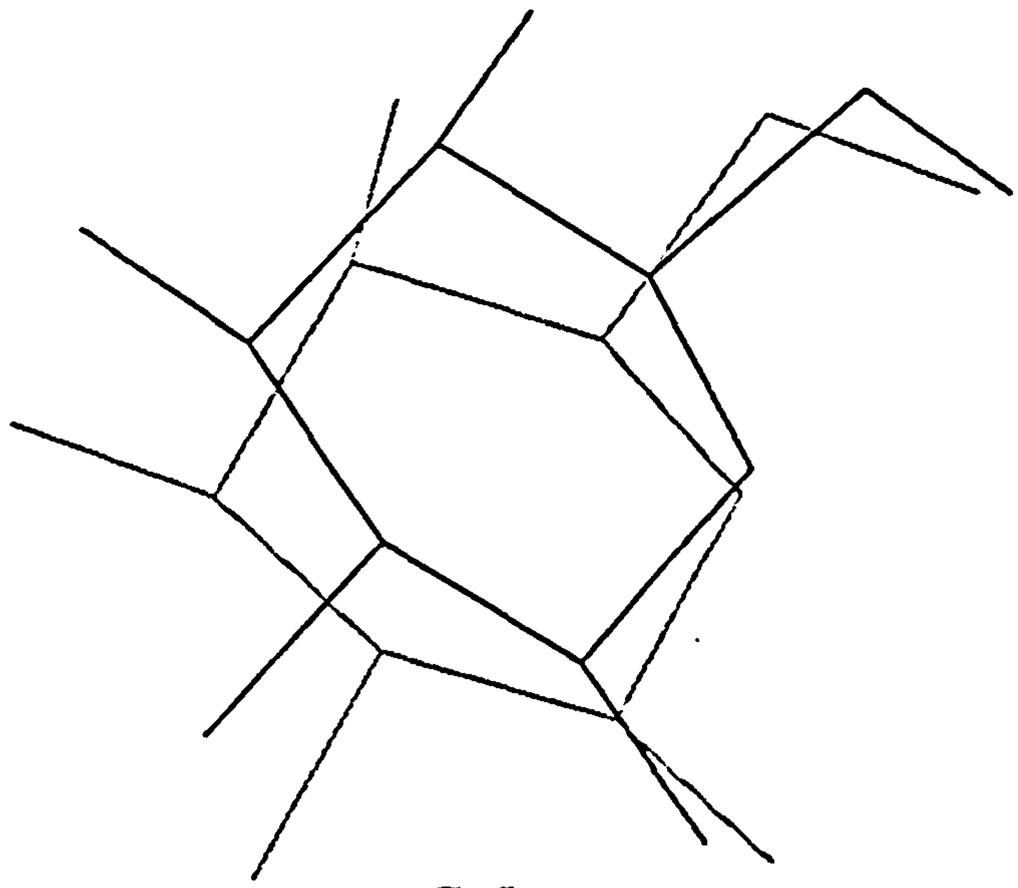


Fig. 4.7 Predictions and experimental conformations for the glycinamide ribonucleotide transformylase complex with Burroughs-Wellcome inhibitor 1476u89.



Gal

Fig. 4.8 Prediction and experimental conformation for the glucose/galactose-binding protein complex with galactose.

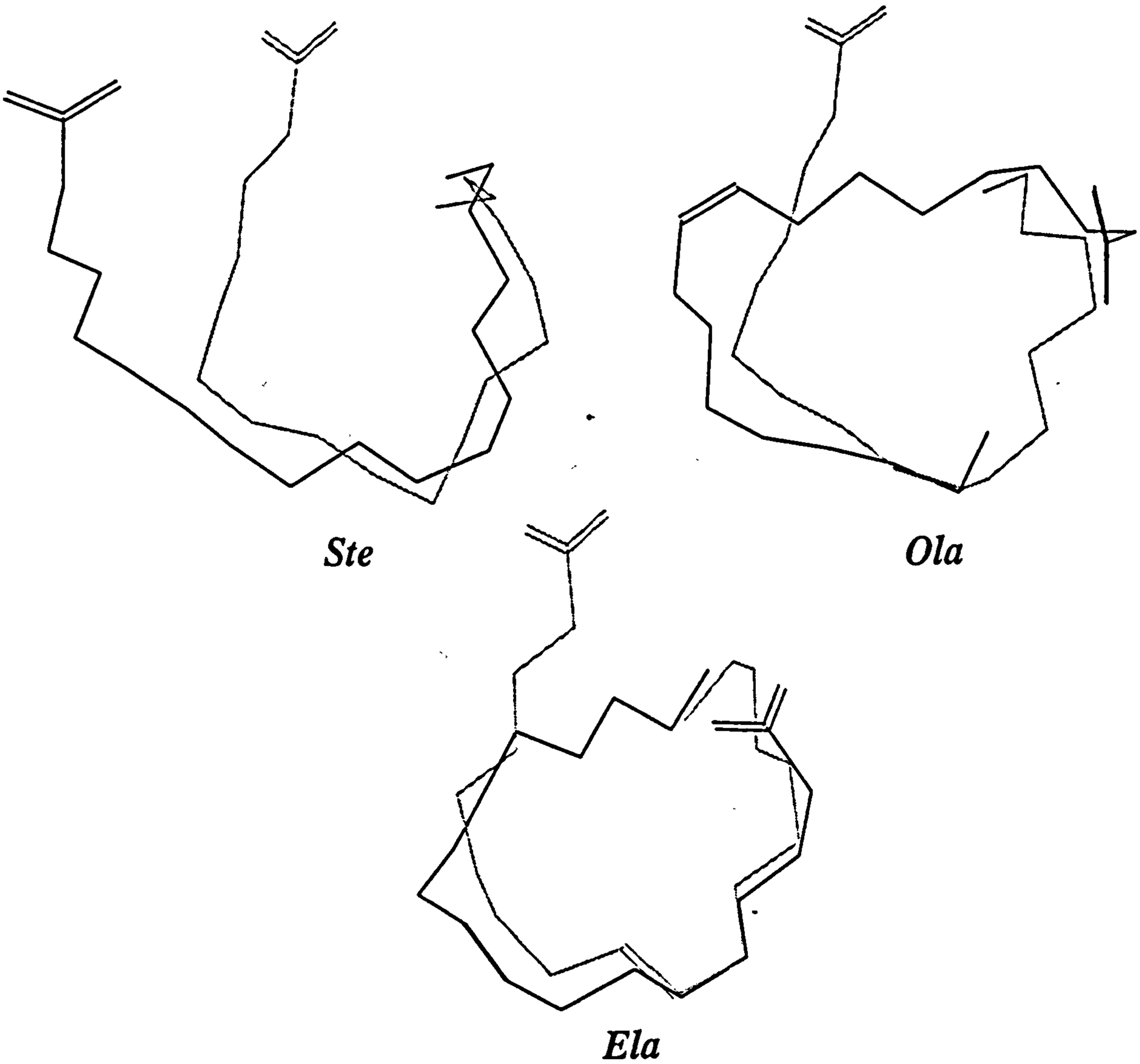


Fig. 4.9 Predictions and experimental conformations for the fatty acid binding protein complexes with elaidic, oleic and stearic acids.

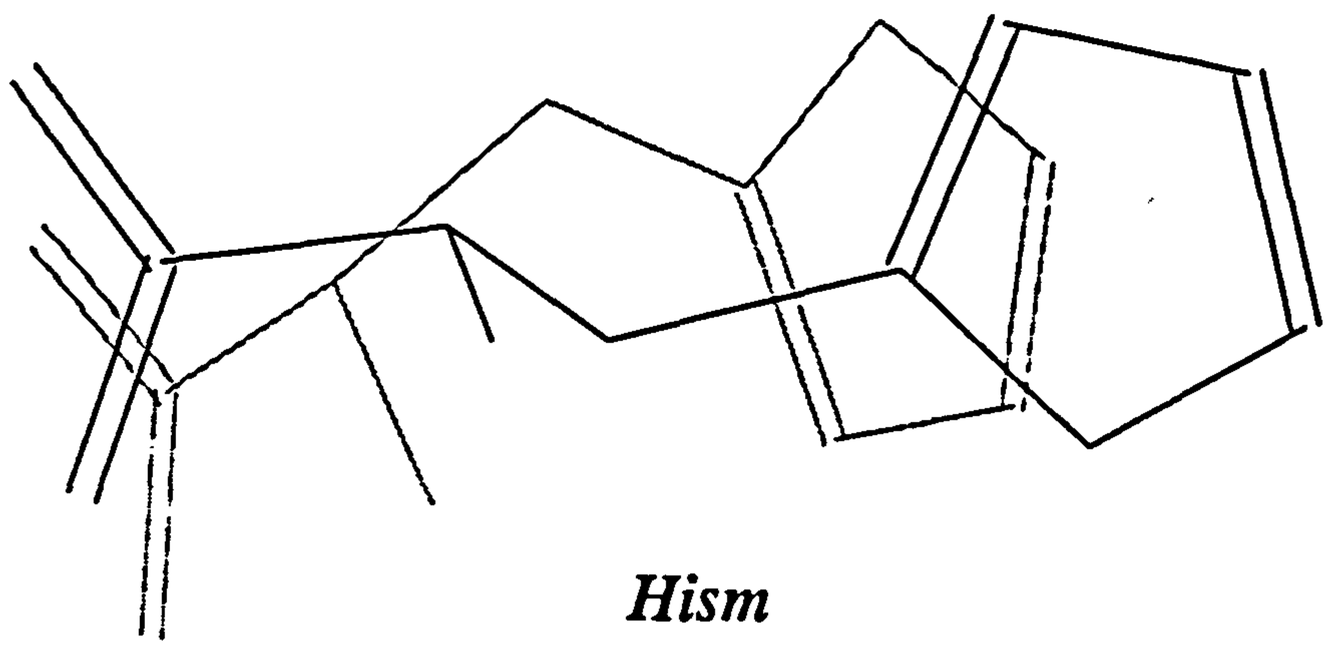


Fig. 4.10 Prediction and experimental conformation for the histidine-binding protein complex with histidine.

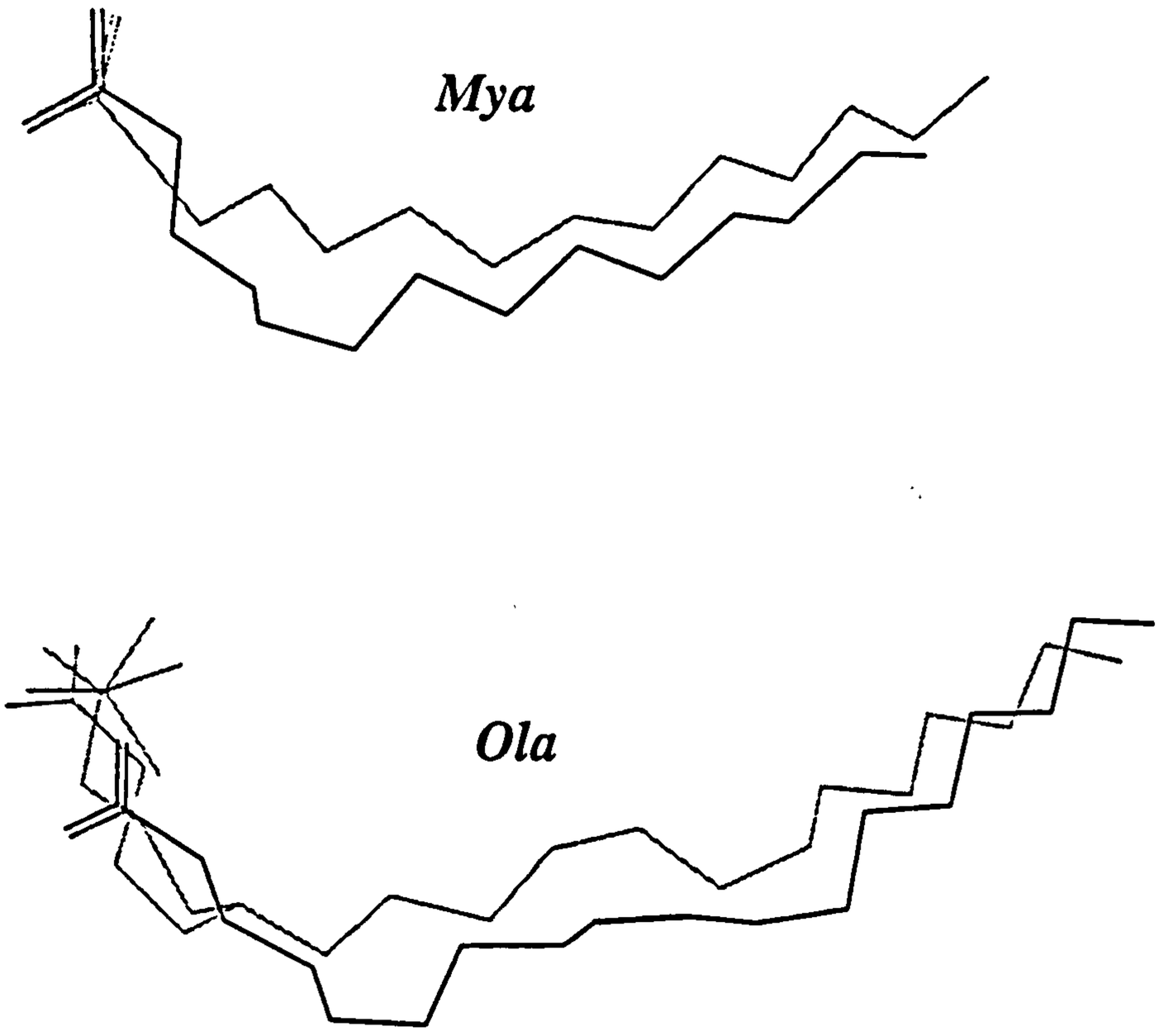


Fig. 4.11 Predictions and experimental conformations for the intestinal fatty acid binding protein complexes with myristate and oleate.

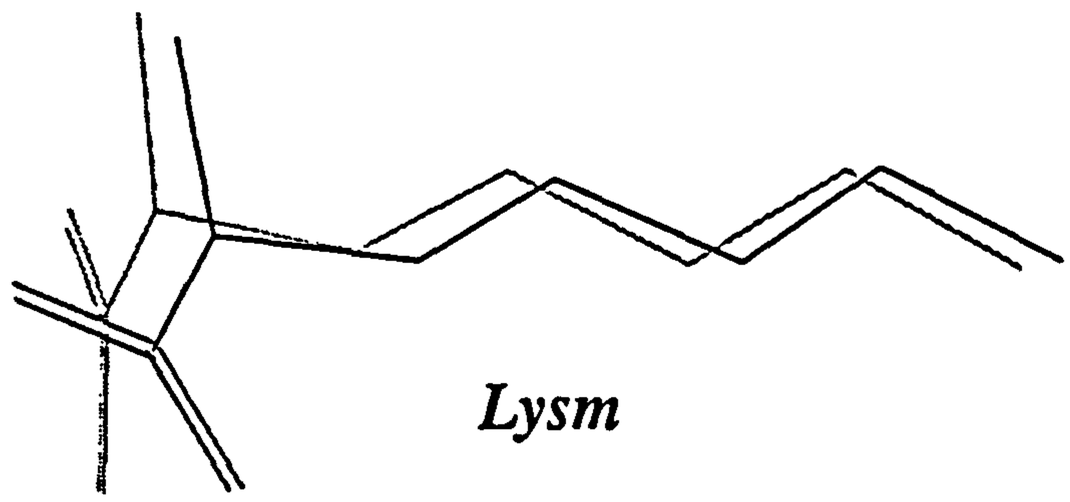


Fig. 4.12 Predictions and experimental conformations for the lysine-, arginine-, ornithine-binding protein complex with lysine.

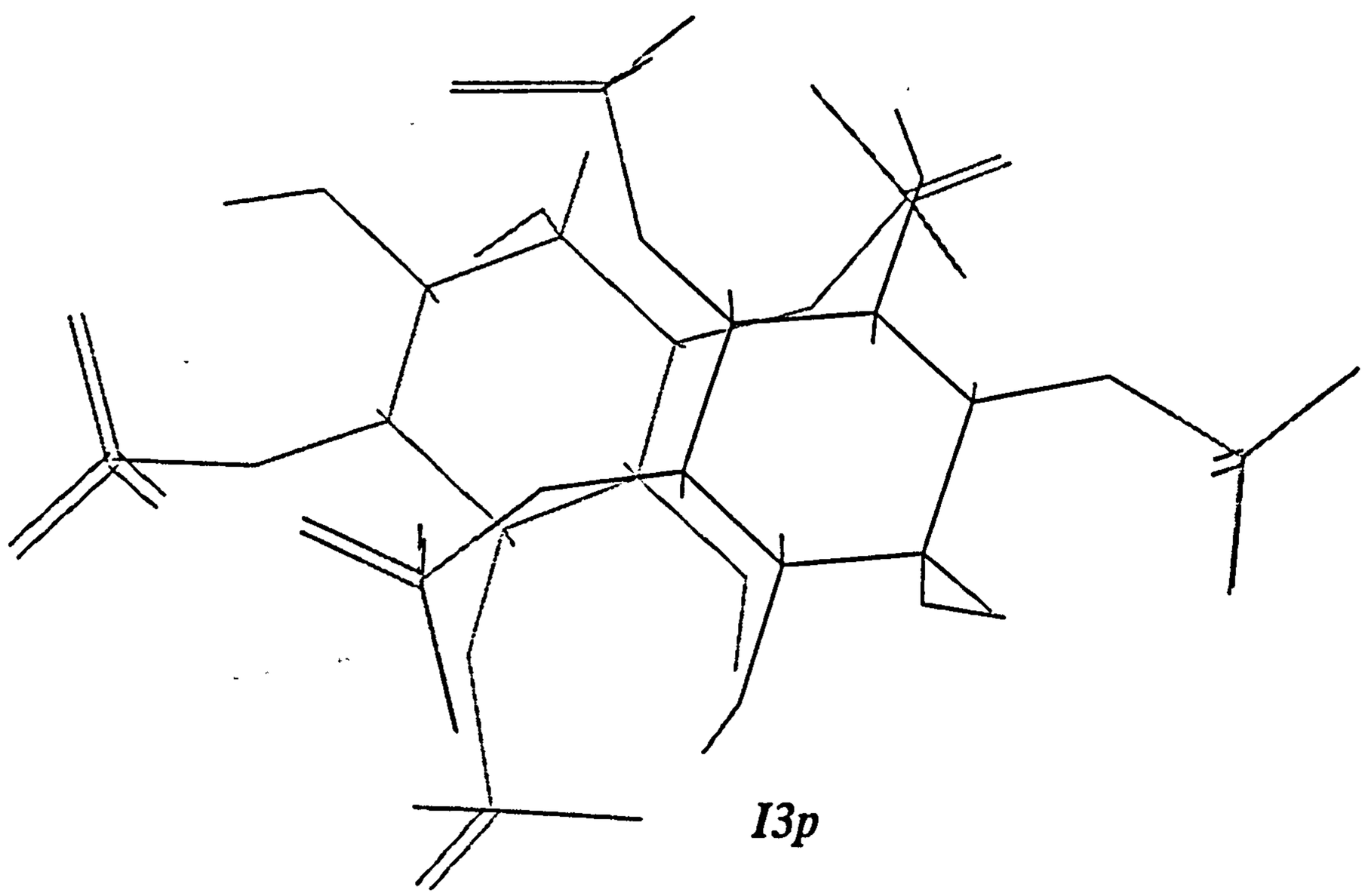


Fig. 4.13 Prediction and experimental conformation for the phospholipase c δ-1 complex with inositol trisphosphate.

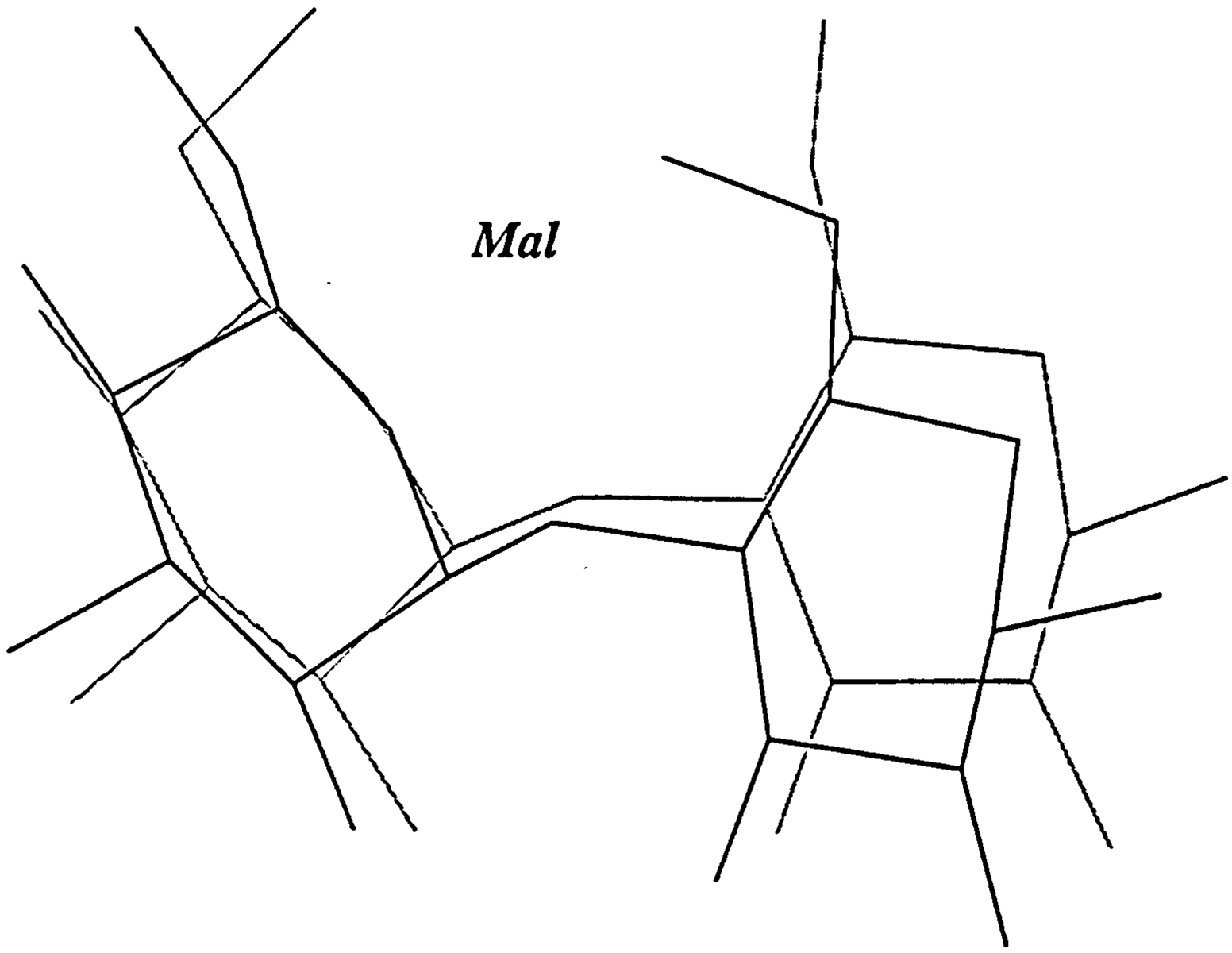
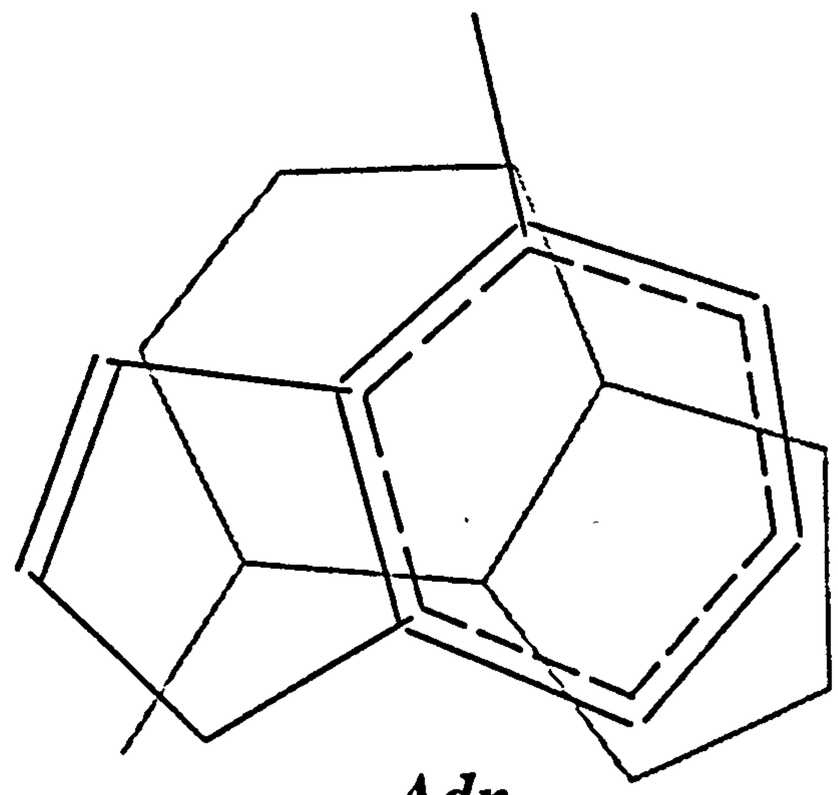


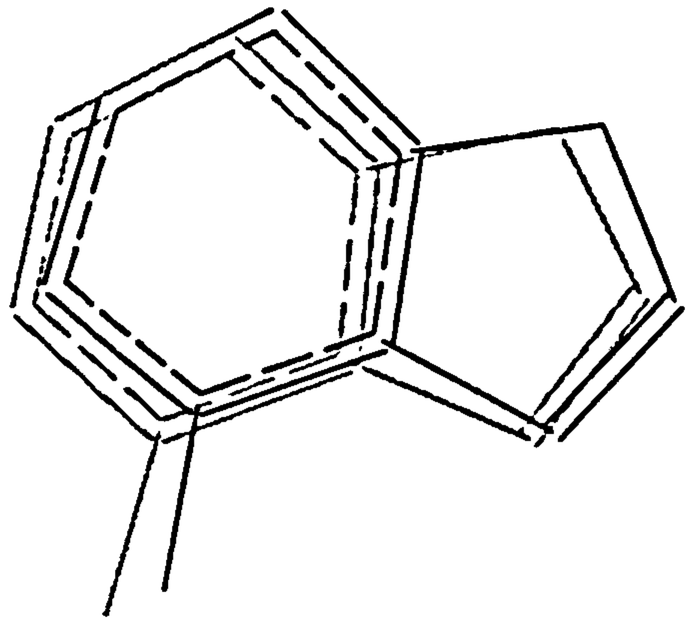
Fig. 4.14 Predictions and experimental conformations for the maltodextrin-binding protein complex with maltose.

α-momorcharin complex with adenine



Adn

Fig. 4.15 Predictions and experimental conformations for the α-momorcharin complex with adenine.



Adn

Fig. 4.16 Prediction and experimental conformation for the α-trichosanthin complex with adenine.

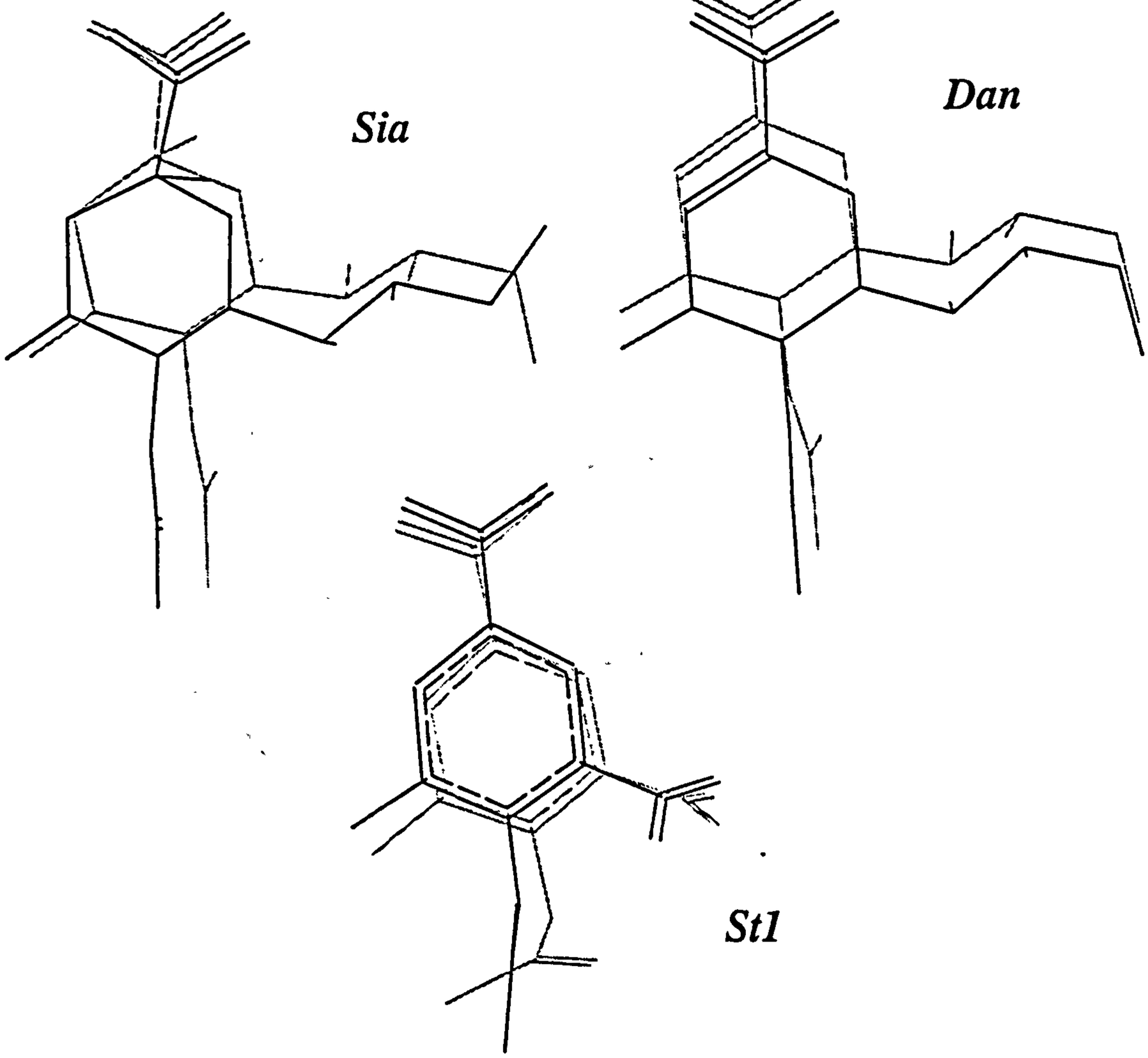


Fig. 4.17 Predictions and experimental conformations for the neuraminidase complexes with sialic acid, 2,3-dehydro-2-deoxy-n-acetyl neuraminic acid and 4-(acetylamino)-3-hydroxy-5-nitrobenzoic acid.

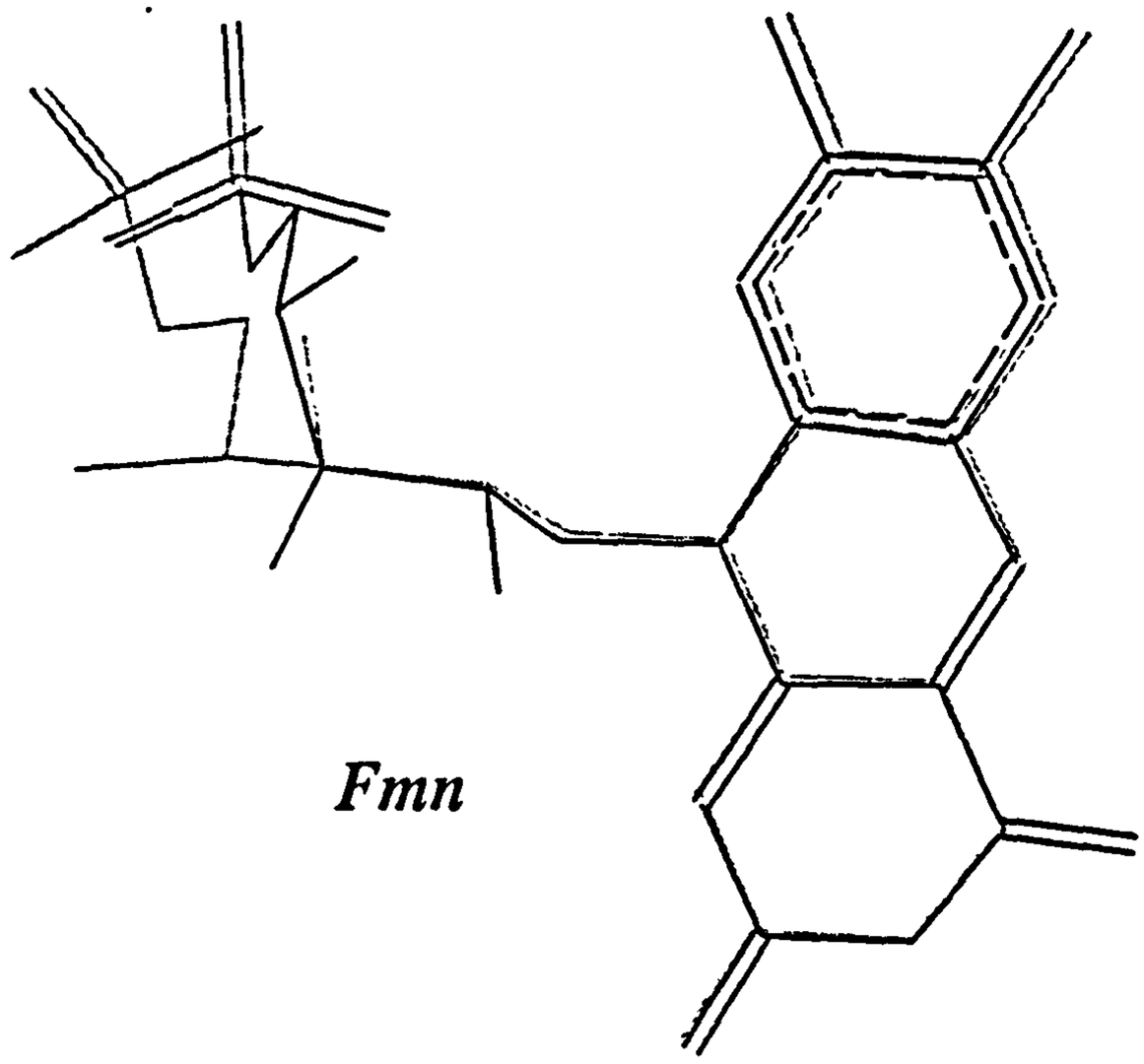


Fig. 4.18 Prediction and experimental conformation for the flavodoxin complex with flavin mononucleotide.

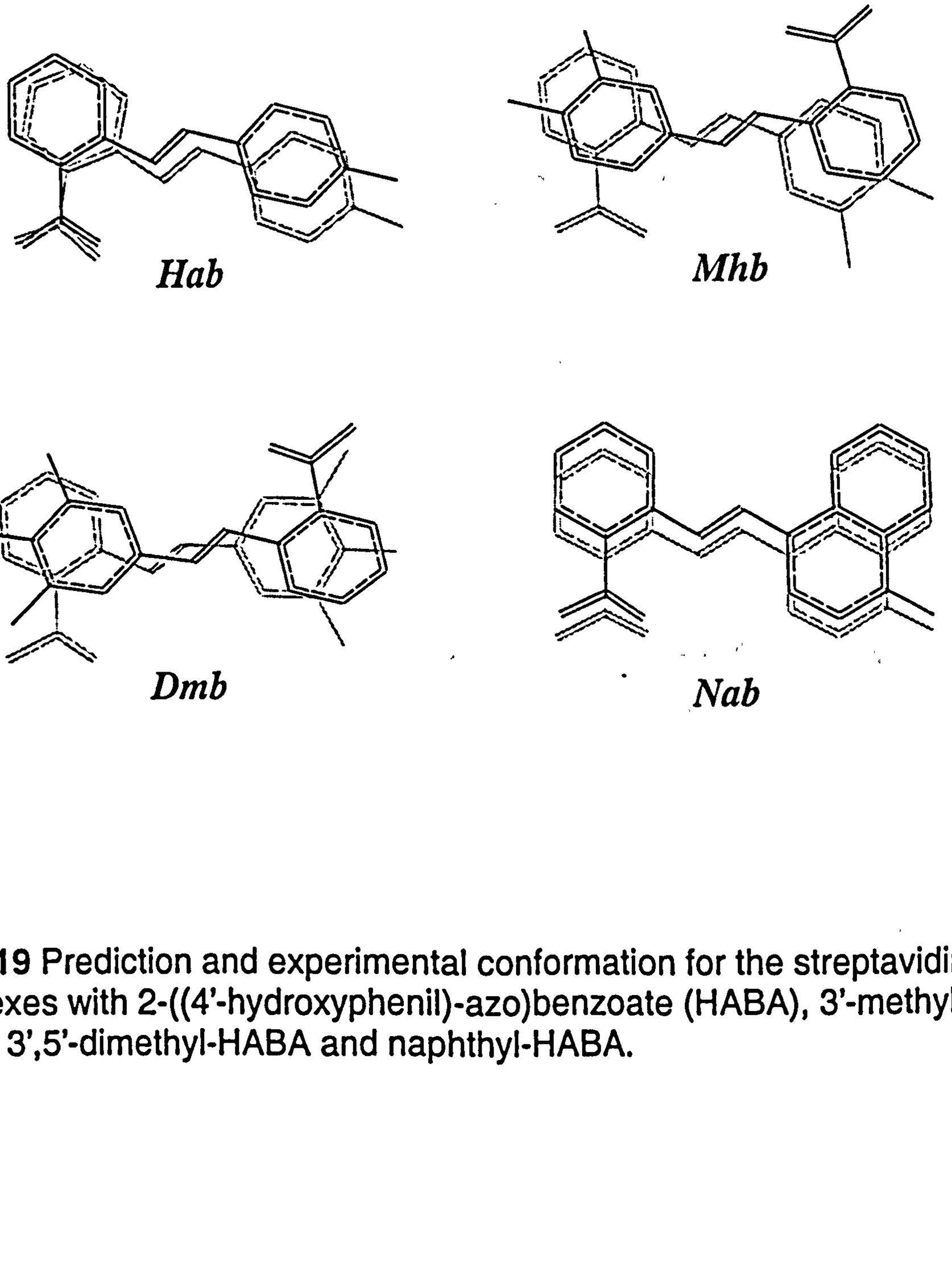
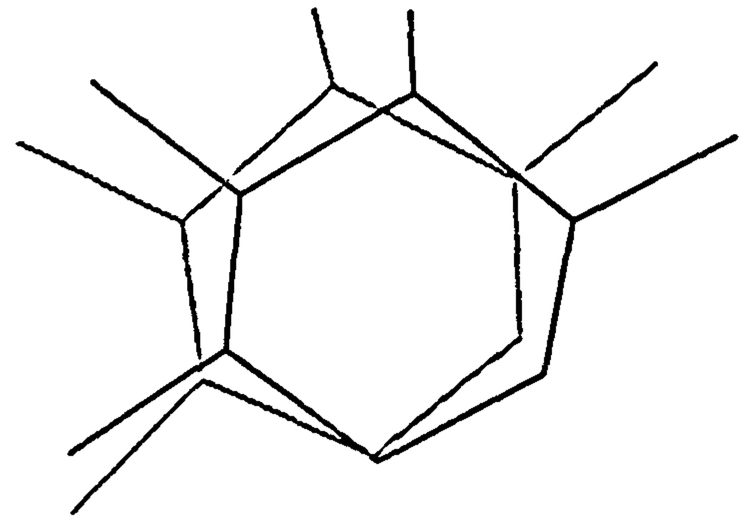


Fig. 4.19 Prediction and experimental conformation for the streptavidin complexes with 2-((4'-hydroxyphenil)-azo)benzoate (HABA), 3'-methyl-HABA, 3',5'-dimethyl-HABA and naphthyl-HABA.



Rip

Fig. 4.20 Predictions and experimental conformations for the D-ribose-binding protein complex with beta-d-ribose.

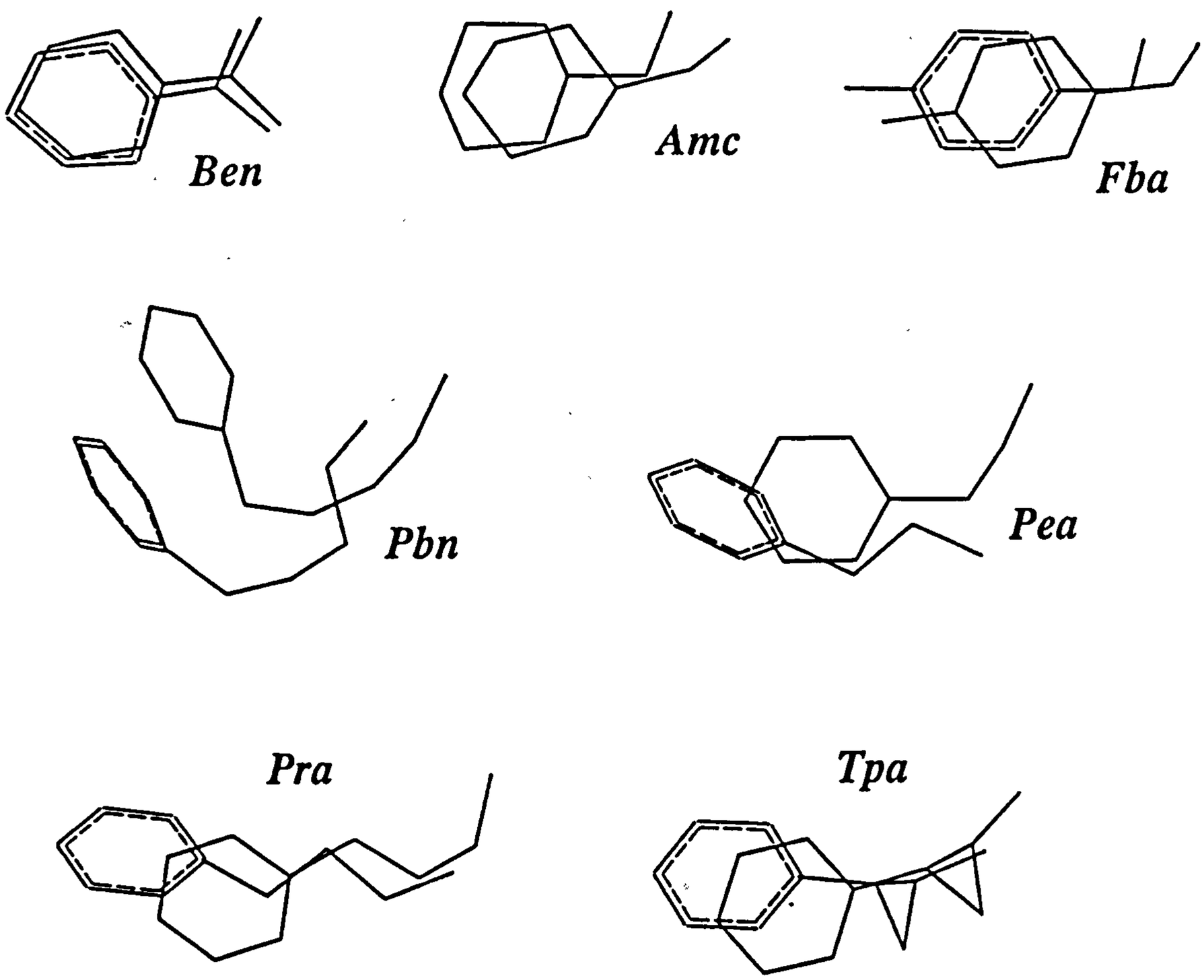


Fig. 4.21 Predictions and experimental conformations for the trypsin complexes with inhibitors benzamidine aminomethylcyclohexane 4-fluorobenzylamine 4-phenylbutylamine 2-phenylethylamine 3-phenylpropylamine tranylcypromine.

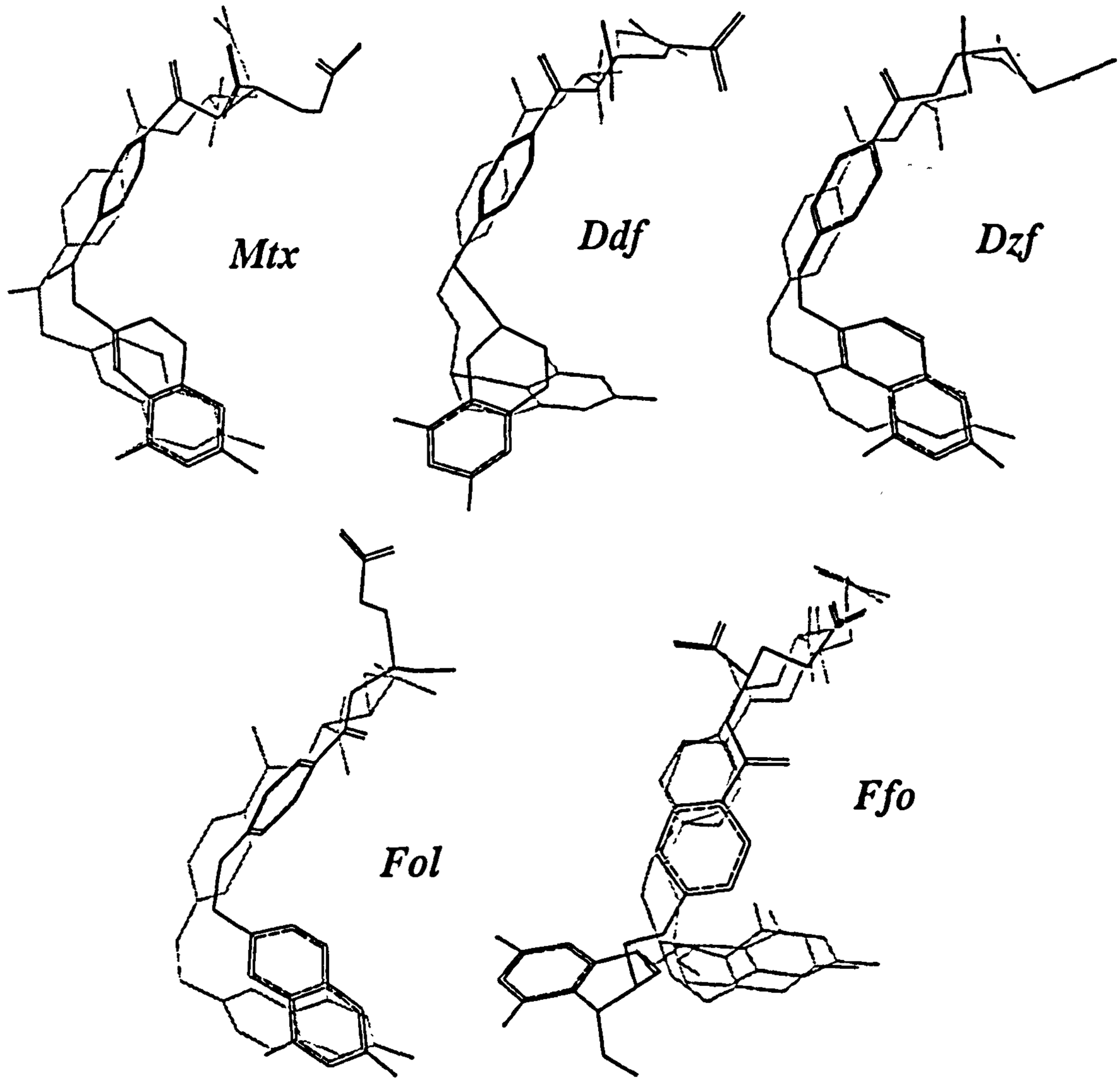


Fig. 4.22 Predictions and experimental conformations for the dihydrofolate reductase complexes with methotrexate 5,10-dideazatetrahydrofolate 5-deazafolate folate folinic acid.

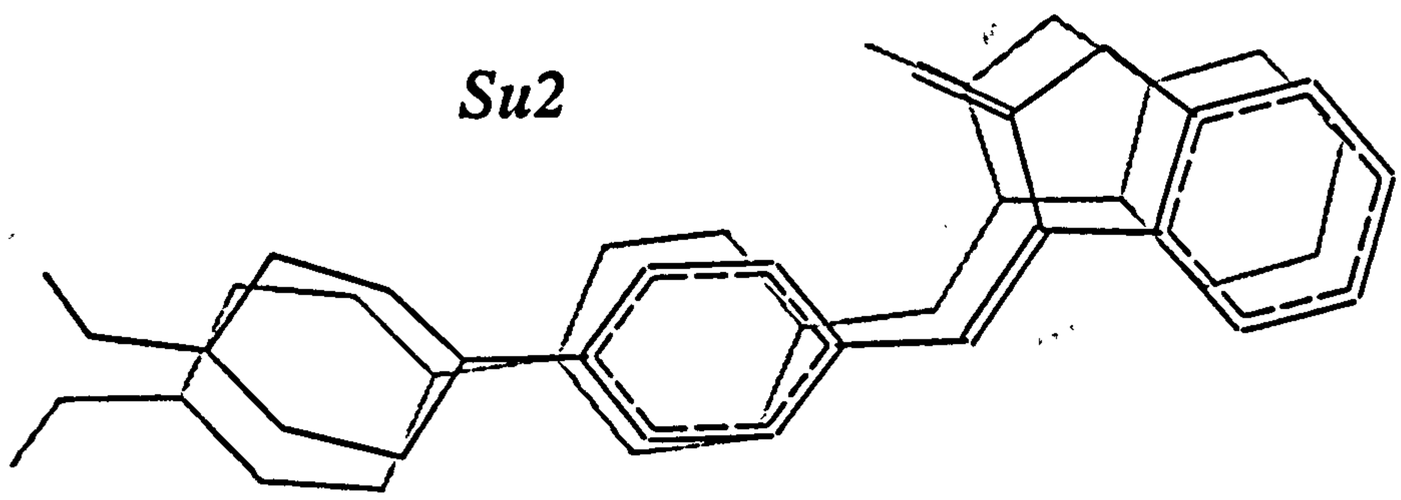


Fig. 4.23 Prediction and experimental conformation for the tyrosine kinase of FGF receptor complex with Sugen inhibitor.

4.3.2 Optimisation of discrimination potential

The majority of the optimisations converged to α=(0.79 1.46 0.44 0.9 1.48). This combination of weights more than doubled the number of correctly recognised ligands, confirming the success of the optimisation. We also tested if any of the terms is superfluous by attempting optimisations with one of the terms excluded. In all such runs we found that the optimised parameter sets showed considerably poorer recognition ability measured by the number of native ligands placed below all false positives. Thus the minimum found is meaningful and there is no redundancy in the set of terms used. As the values of weights show, the hydrophobic and hydrogen bonding contributions are more important than we assumed initially, suggesting the cost of one ideal hydrogen bond is almost 4 kcal/mol and the surface tension is close to 45 cal/mol/Å^2. There was a more than two-fold decrease in the weighting of the solvation electrostatics contribution while the distance-dependent term achieved considerable weight in the optimised function. This possibly indicates that the solute-solvent part of the solvation electrostatics might be overestimated when calculated with inner dielectric constant ε=4. The resulting distributions of the discrimination potential for all ligands and receptors are shown on Fig. 4.24

Interestingly, in most cases optimised potential would assign to the native ligands the energy values close to -30 kcal/mol. Though we did not directly attempt to obtain an absolute measure of binding affinity, this observation suggests that with a constant of around 25 kcal/mol added, the optimised potential might be used as an approximation of binding free energy.

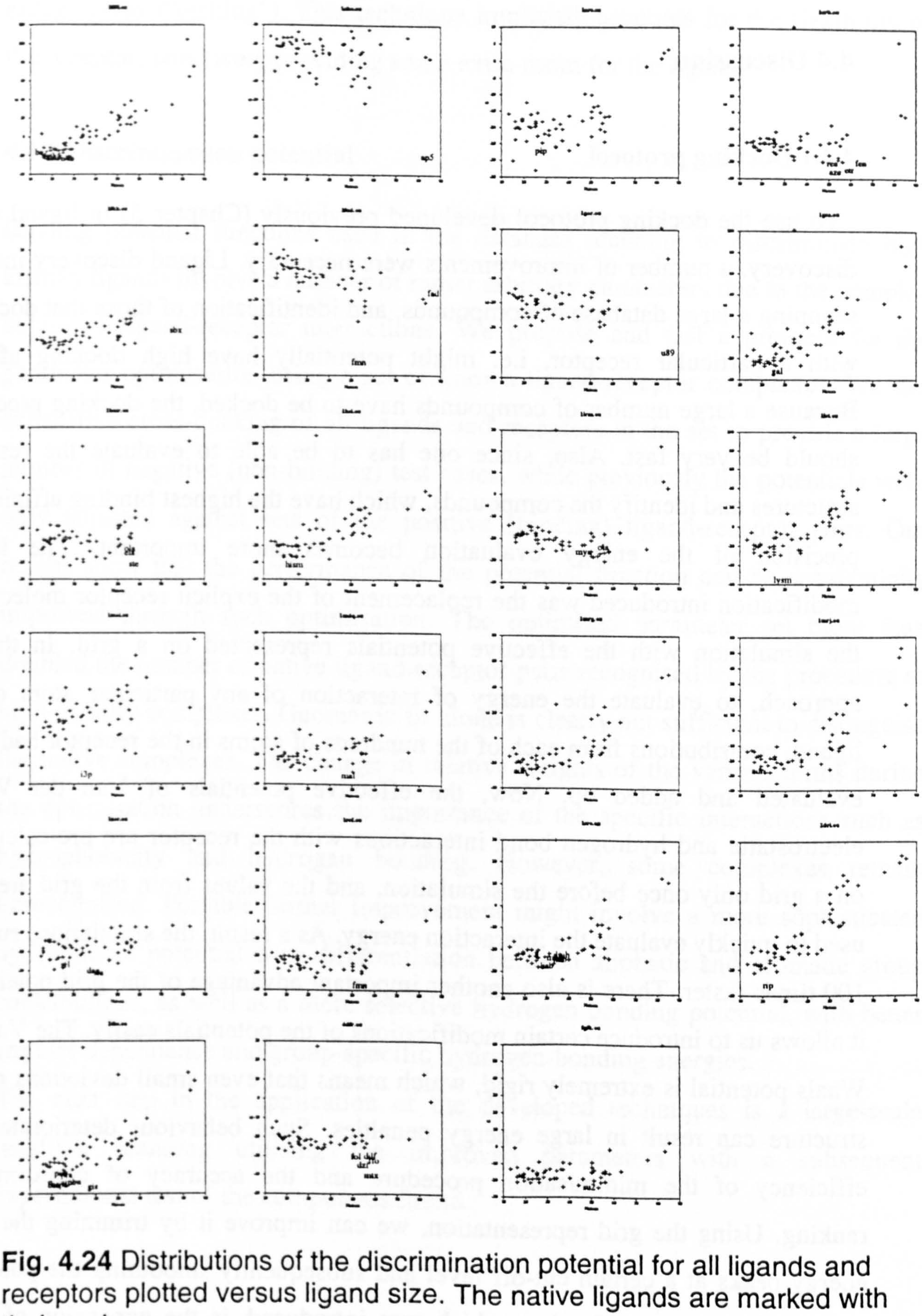


Fig. 4.24 Distributions of the discrimination potential for all ligands and receptors plotted versus ligand size. The native ligands are marked with their code-names.

4.4 Discussion

4.4.1 Docking protocol

To use the docking protocol developed previously (Chapter 3) in ligand (drug) discovery, a number of improvements were necessary. Ligand discovery involves scanning a large database of compounds, and identification of those that dock best with a particular receptor, i.e. might potentially have high docking affinity. Because a large number of compounds have to be docked, the docking procedure should be very fast. Also, since one has to be able to evaluate the resulting structures and identify the compounds, which have the highest binding affinity, the precision of the energy evaluation becomes more important. The largest modification introduced was the replacement of the explicit receptor molecule in the simulation with the effective potentials represented on a grid. In the old approach, to evaluate the energy of interaction of any particular atom of the ligand, contributions from each of the hundreds of atoms in the receptor had to be evaluated and added up. Now, the effective potentials of Van der Waals, electrostatic and hydrogen bond interactions with the receptor are pre-calculated on a grid only once before the simulation, and the values from the grid are later used to quickly evaluate the interaction energy. As a result, the simulations run 50-100 times faster. There is also another important advantage of the grid potential - it allows us to introduce certain modifications of the potentials easily. The Van der Waals potential is extremely rigid, which means that even small deviations of the structure can result in large energy penalties. Such behaviour deteriorates the efficiency of the minimisation procedure and the accuracy of conformation ranking. Using the grid representation, we can improve it by trimming the high energy peaks at a certain cut-off level and subsequently smoothing the potential surface. Another modification, which was introduced, is the expansion of low-

energy areas ("etching"). This technique implicitly accounts for the flexibility of the receptor, effectively providing some extra room for the ligand.

4.4.2 Discrimination potential

Binding potential functions used in the database scanning to discriminate high affinity ligands involve a number of rather arbitrary parameters due to the complex nature of ligand-receptor interactions. We propose and test a protocol for the parameter optimisation using a set of known ligand-receptor complexes. We use exhaustive cross-docking of all ligands and receptors in the set to provide a large number of negative (non-binding) test cases, while previously the potentials were only adjusted against sets of the positive (binding) ligand-receptor pairs. Our results show that the performance of the potential function can be substantially improved through such optimisation. The optimised parameter set more than doubled the number of native ligand-receptor pairs recognised by the procedure as high-affinity complexes. Geometric fit alone is clearly not sufficient to distinguish the native complexes. The change in relative weights of the various terms during the optimisation underscores the importance of the specific interactions such as hydrophobicity and hydrogen bonding. However, some complexes remain unrecognised. Possible further improvement might involve a more sophisticated hydrophobic potential with differentiation between aliphatic and aromatic group contributions, as well as a more selective hydrogen bonding potential, with better angular dependence and group-specific hydrogen-bonding energies.

The next step in the application of the developed techniques is a large-scale database scanning utilising the improved parameters with a subsequent experimental test of the compounds found.

5. Ligand discovery by virtual database screening: novel ligands of FGFR.

5.1 Introduction

5.1.1 Chemical database screening

One of the most important applications of docking simulations is the discovery of drug leads, i.e. novel compounds that bind to a particular receptor. Previously, computational methods for the identification of the potential new ligands were limited to chemical similarity scanning. Such methods require the knowledge of other ligands beforehand, can hardly identify any substantially novel compounds and dismiss the essentially three-dimensional character of the protein-ligand interactions. As the techniques of protein structure determination mature, the three-dimensional structures of the receptors and enzymes implicated in many pathological processes become readily accessible and can be used to search large databases of commercially or synthetically available chemical compounds using novel docking techniques. A number of attempts at structure-based drug discovery have been reported in the past 5 years. Shoichet at al. [92] scanned the Fine Chemicals Directory of 55,000 compounds for the inhibitors of thymidylate synthase using program DOCK [93], which identified 600 compounds from which 25 candidates were selected manually. 3 of them have shown some activity, albeit rather weak (inhibition constant in the high micromolar range). Crystal structure was solved for one of the complexes and the binding mode was found to be different from the one predicted by DOCK, making this work only a partial success. More recently, Hoffman et al. [94] reported successful identification of low-micromolar inducer of the conformational change in the influenza virus hemagglutinin using an improved DOCK program to scan approximately 150,000

compounds. However, no experimental confirmation of the structure of the complex was attempted.

The application of docking to the database screening imposes new requirements on the docking procedure. Since the number of compounds to be docked can be very large, the speed of the docking routine becomes more important than in the case where only a single ligand is considered. Accurate evaluation of the binding affinity becomes crucial for the selection of a few candidates to be tested experimentally out of many thousands.

5.1.2 Tyrosine kinases and fibroblast growth factor receptor

Tyrosine kinases (TKs) are important components of signalling pathways controlling cell proliferation and differentiation. Many cell membrane receptors contain tyrosine kinase domains in their intracellular part which self- or cross-phosphorylate in response to the binding of various factors to the extracellular part. Fibroblast Growth Factor Receptor (FGFR) plays an important role in embryonic development, angiogenesis, wound healing and malignant transformation [95]. Improper expression or activation of this receptor has been implicated in several skeletal disorders and angiogenic pathologies (e.g. [96]). Amplification and overexpression of FGFR has been detected in a number of cancers (e.g. [97]). The structure of the TK domain of FGFR was solved by X-ray crystallography [98].

Since the initial discovery of the TKs and their role in signal transduction and growth regulation [99,100], investigators soon began looking for selective inhibitors of various TKs, and the area is a subject of intense research (reviewed in [101]).

Here we apply previously developed (Chapters 3 and 4) docking and ligand discrimination techniques to the specific case of database screening for the inhibitors of the tyrosine kinase of the fibroblast growth factor receptor.

5.2 Materials and Methods

5.2.1 Drug-likeness filtering

Apart from the binding affinity and specificity, there are a number of other requirements for drug candidates, such as bioavailability and low toxicity. Bioavailability is the capability of the drug to attain its target in the organism after being administered. The drug typically has to cross several barriers in the organism, especially in the case of oral administration, which is generally the preferred mode. A number of empirical criteria have been established [102]. In our pre-selection algorithm we utilised the following rules:

-molecular weight of the selected compounds was between 100 and 500 Dalton,

-the number of hydrogen bond acceptors did not exceed 10,

-the number of hydrogen bond donors did not exceed 5.

Very small ligands rarely exhibit sufficient specificity because of the very limited number of interactions with the receptor. Large compounds are more likely to have delivery problems and have limited optimisation potential. Excessive hydrogen-bonding capacity interferes with successful membrane crossing, since such compounds often have a very unfavourable partitioning coefficient between water and lipids.

Water solubility is essential for most drugs and in the case of particular study it was important to make the experimental *in vitro* tests possible. Insoluble compounds often display otherwise good binding properties because their hydrophobicity promotes complexation. Thus, screening protocol generally favours such compounds and it is important to filter them out if possible. We used an estimate of solvation energy as a solubility criterion in one of the filters. Solvation energy was calculated as a sum of the electrostatic solvation energy calculated by boundary element method with an internal dielectric constant of 8,

and of surface tension with the constant of 8.23 cal/mole/Å. Compounds with the values above 1.8 kcal/mole were discarded. The threshold represents a crude estimate of the entropic energy gain of $6*(1/2)k_BT \approx 3*0.6$ kcal/mole when 6 degrees of freedom (three translations and three rotations) are added for each molecule upon its dissolution.

5.2.2 Docking

If the ligand satisfied the pre-screening criteria, it was subjected to the docking procedure. The protocol used in this work was a modified version of the protocol developed in the previous stage of this project (Chapter 4). To reduce further the time required for the docking of each individual compound, the single Monte-Carlo simulation was replaced by a "map-annealing" procedure described below.

The basic idea of map annealing is to dock the ligand initially into a very simplified and smoothed image of the binding site and then gradually adjust the solutions to the more and more exact versions of it. Such a protocol should quickly find a rough solution and then would only need to introduce minor improvements as the approximation of the receptor potential becomes more exact. Practically, several sub-optimal solutions have to be kept since in many cases best low-resolution conformation is not in the vicinity of the actual answer. Fortunately, the ICM conformational stack (see sec. 2.1.1) is ideally suited for keeping such a set of conformations. Practically, the protocol consisted of

- preparation of 3 sets of potential maps, the final set and two other sets smoothened by iterative application of the formula $P^{i+1}_{j,k,l}=((P^i_{j-1,k,l} + P^i_{j+1,k,l} + P^i_{j,k+1,l} + P^i_{j,k-1,l} + P^i_{j,k,l+1} + P^i_{j,k,l-1})/6 + P^i_{j,k,l})/2$, which is a simple discrete implementation of spatial averaging of the potentials. The most approximate set of maps was produced after 20 iterations and the intermediate after 5.
- Putative ligand is subjected to a short Monte-Carlo run without any presence of the receptor. Up to 50 low energy conformations are accumulated in the

stack during the run. To ensure diversity in this set only the conformations with torsion angular RMSD higher then 45° are retained. All conformers are then translated into an arbitrarily chosen initial position in the binding pocket.

- Each conformer in the stack is duplicated and rotated 180° around its short axis to facilitate the sampling.
- Each conformer is subjected to local minimisation in the set 3 of the potential maps. The stack is sorted according to the approximate interaction energy and the best conformer is subjected to a Monte-Carlo minimisation. During the MC run, some of the new conformations may get stored in the stack or replace old ones according to the stack maturation protocol (described in sec. 2.1.1). If no new or better conformation is found within 10 MC steps, the procedure switches to another stack conformation (stack jump). Stack jump allows Monte-Carlo improvement of conformers other than the best one.
- Same as above, but with the potential maps of set 2.
- Same as above, but with the potential maps of set 1.
- Best conformer is taken as the final solution. To produce a model without significant steric clash, the ligand is placed into rigid full-atom model of the receptor and subjected to local energy minimisation with weak harmonic restraints to the original solution.

The discrimination potential was then calculated for the final docking solution to make a decision if the given compound should be stored or discarded.

5.2.3 Discrimination potential

The potential function used was derived as described in chapter 4 to discriminate putative high-affinity ligands. It included electrostatic, hydrogen-bonding, hydrophobic, Van der Waals and entropic contributions. To establish the threshold of the potential, a preliminary scan of an arbitrary subset of the database (3000

compounds) was used, as well as an evaluation of the potential for a known inhibitor with solved 3D structure of the complex, SU4984. The latter gave the discrimination potential value of –26.1 . Some of the compounds in the subset gave values of the potential as low as –30.0 (see Fig. 5.1). To retain candidates with an affinity similar to the known compound and to keep a reasonable number of compounds, the threshold of -25.5 was chosen.

5.2.4 Experimental tests

To confirm the ability of the developed scanning protocol to identify binding ligands we have tested the binding ability of the compounds from the final selection list in a direct experimental assay. The assay was based on the autophosphorylation ability of the FGFR TK domain. Free FGFR TK in presence of ATP quickly phosphorylates a number of tyrosine residues on its surface, changing considerably the total electric charge of the molecule. The presence of a ligand bound in the active site inhibits the enzymatic activity. Thus, depending on the binding affinity of the ligand, incubation of FGFR TK with ligand and ATP results in varied degrees of the phosphorylation. The latter can be subsequently measured by gel electrophoresis which is sensitive to the difference in the electric charge of the phosphorylated and unphosphorylated protein.

FGFR TK was incubated with 500μM ATP and 500μM putative ligand solution.

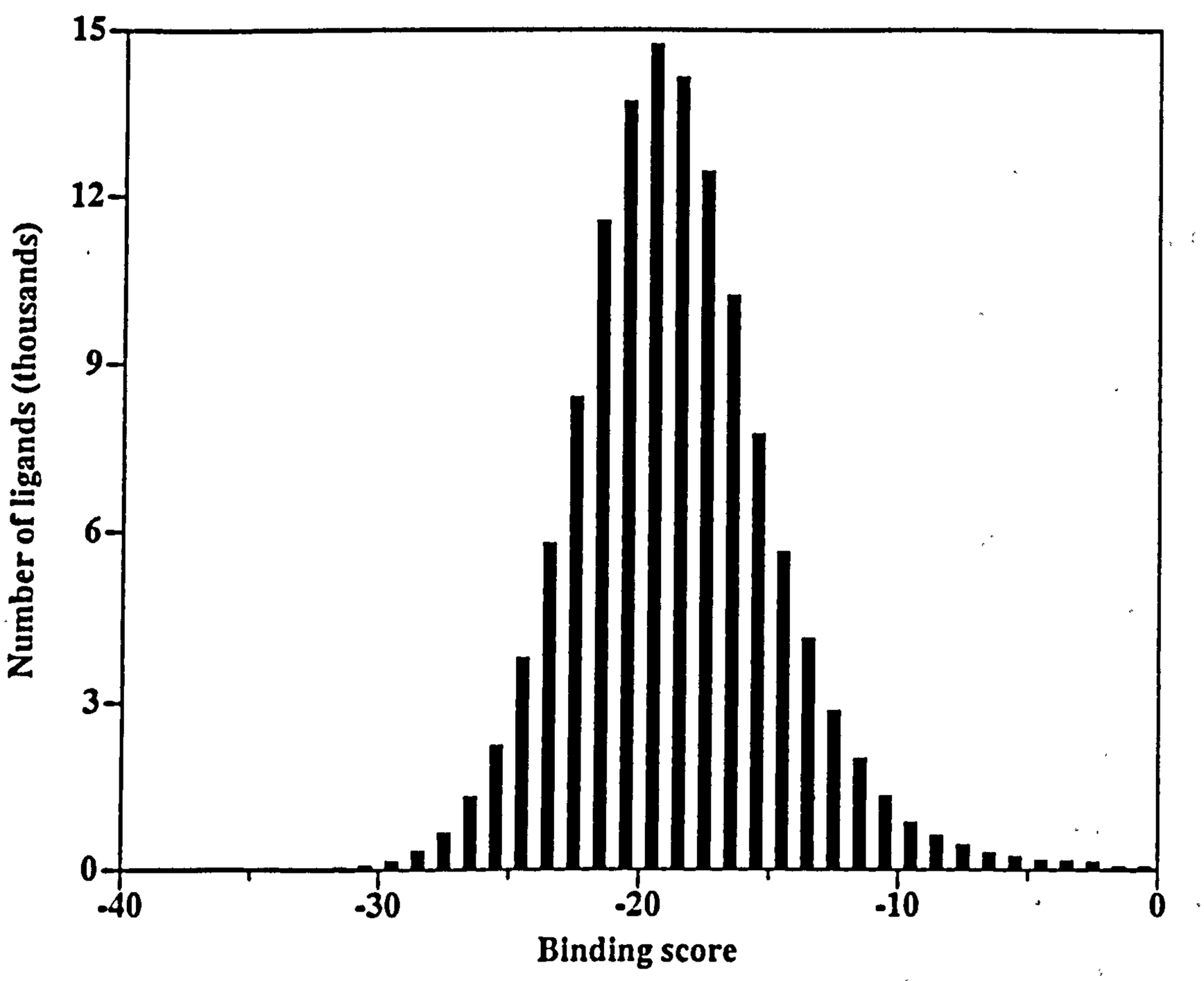


Fig. 5.1 The histogram for the distribution of the discrimination potential values for the entire database

5.3 Results

5.3.1 Virtual Screening

153000 compounds in the ACD library were scanned. About 25% of the compounds were discarded by the simple drug-likeness criteria. The main screening procedure selected and stored predictions of the bound conformation for each of 3821 ligands which had the discrimination potential below the cutoff of -25.5 . Calculations were performed on a 20-CPU symmetric multiprocessing machine, SGI Onyx R10000. To accelerate scanning, a simple "database striping" parallelisation technique was used, dividing the database into 15 subsets of approximately 10,000 compounds each, which were scanned in parallel on 15 CPUs, taking about 240 hours to complete the scan. The histogram for the distribution of the discrimination potential for the entire database is plotted on Fig. 5.1. To illustrate the docking ability of the procedure, 100 docked ligands are shown inside the binding pocket of the receptor on fig. 5.2.

The bound conformations of selected compounds were than used in the further filtering with the hydrogen bond formation criterion. Examination of several known structures of the FGFR tyrosine kinase complexes has shown that formation of two hydrogen bonds might be particularly important for ligand binding, one to the carboxyl oxygen of glutamic acid residue 562 and another to the amido hydrogen of alanine residue 564. Only ligands which had a polar hydrogen atom within 2.5Å from the carboxyl oxygen and a hydrogen bond acceptor atom within the same cutoff distance from the amido hydrogen were retained. This simple hydrogen-bond formation criterion yielded 185 compounds. We calculated an estimate of the solvation energy for all of them and dismissed a further 23 compounds with the value above –1.8 kcal/mole as described above. The final list contained 162 compounds.

5.3.2 Experimental tests

Out of 162 potential ligands selected in the last stages of virtual screening, from practical considerations we purchased and tested 53 compounds available from Maybridge Chemical Company, Sigma-Aldrich library of rare compounds and Sigma Chemical Company catalogue. Some manufacturers were difficult to locate and contact or only had 1 or 2 compounds on our list, several compounds were known to be highly poisonous and a considerable number of compounds were either discontinued or not in stock. Only 4 of the purchased compounds had poor solubility, which confirmed the adequacy of the solvation energy prediction filtering. 49 compounds (see Table 5.1) were tested in the phosphorylation inhibition assay, and 5 of them showed activity. The gel electrophoresis results can be seen on Fig. 5.3 (a-c). Predicted bound conformations for the active compounds are shown on Fig. 5.4 (a-e).

Table 5.1 Compounds selected for the experimental inhibition activity tests.

Some of the compounds listed were not tested for the reasons marked in the "comments" column.

N	Index in the database	Manufacturer and catalogue number	Binding score	Comments
1	83384	Maybridge btb_05310	-27.39	
2	83653	Maybridge btb_05838	-27.61	
3	71589	Maybridge btb_06100	-25.83	
4	71588	Maybridge btb_06212	-25.61	
5	84074	Maybridge btb_07226	-26.31	
6	74513	Maybridge btb_08002	-26.79	
7	76799	Maybridge btb_08046	-26.11	
8	123369	Maybridge btb_08301	-25.90	
9	47013	Maybridge btb_08390	-26.672	
10	67642	Maybridge btb_09751	-30.07	Not received
11	124904	Maybridge btb_09809	-28.39	
12	81504	Maybridge btb_10623	-25.97	
13	81507	Maybridge btb_10631	-25.73	**Active**
14	81485	Maybridge btb_10632	-25.93	
15	84253	Maybridge btbt_00050	-26.21	
16	84500	Maybridge cd_00745	-26.35	
17	64660	Maybridge cd_01222	-27.98	
18	85356	Maybridge cd_02746	-26.31	
19	85919	Maybridge cd_04622	-27.06	
20	86951	Maybridge cd_08948	-25.54	
21	87093	Maybridge cd_09631	-26.69	**Active**
22	72307	Maybridge cd_10166	-25.79	
23	87424	Maybridge cd_12001	-26.58	Not received
24	87500	Maybridge dfp_00333	-27.08	Not soluble
25	124257	Maybridge gk_02307	-26.79	
26	81606	Maybridge han_00440	-30.05	Not received

27	150767	Maybridge jfd_00130	-26.13	
28	73181	Maybridge km_03091	-25.52	Not received
29	91947	Maybridge km_04759	-25.72	
30	65216	Maybridge km_06760	-27.97	
31	19773	Maybridge km_06915	-26.56	**Active**
32	124927	Maybridge km_07674	-28.19	
33	146366	Maybridge km_07869	-27.45	
34	68249	Maybridge nrb_01468	-25.98	
35	107222	Maybridge nrb_04029	-28.24	
36	13894	Maybridge nrb_04902	-28.44	
37	148120	Maybridge rjc_00213	-26.43	
38	72981	Maybridge rjf_01361	-25.98	
39	92029	Maybridge sew_04081	-27.10	**Active**
40	81403	Maybridge spb_01635	-28.65	
41	123086	Maybridge spb_02037	-25.90	
42	47294	Maybridge spb_03282	-25.65	
43	145929	Maybridge spb_05775	-26.57	
44	151416	Maybridge spb_06122	-25.61	
45	146817	Maybridge spb_06890	-27.40	
46	34017	Sigma A7783 2'-azido-2'-deoxycytidine	-25.53	Not received
47	11177	Sigma H8502 3-hydroxytyramine hydrochloride	-25.93	
48	6452	Sigma N7389 3-nitro-l-tyrosine	-25.61	Not soluble
49	36052	Sigma A5525 5'-amino-5'-deoxythymidine	-32.13	
50	51533	Sigma E9386 5-ethyl-2'-deoxyuridine	-26.28	
51	4538	Sigma P7644 9-phenyl-2,3,7-trihydroxy-6-fluorone	-26.48	Not soluble
52	94060	Sigma A1437 altertoxin i	-26.96	Poisonous

53	94091	Sigma A1311 austdiol	-27.05	Poisonous
54	15669	Sigma C6007 chrysarobin	-27.16	
55	141111	Sigma L7512 lacmoid	-25.99	**Active**
56	9613	Sigma A0156 n-(4-aminobutyl)-n-ethylisoluminol	-30.09	
57	44825	Sigma A1661 n-(6-aminohexyl)-n-ethylisoluminol	-26.02	
58	137451	Salor s16,362-7	-29.67	
59	127340	Salor s2,949-2	-28.15	

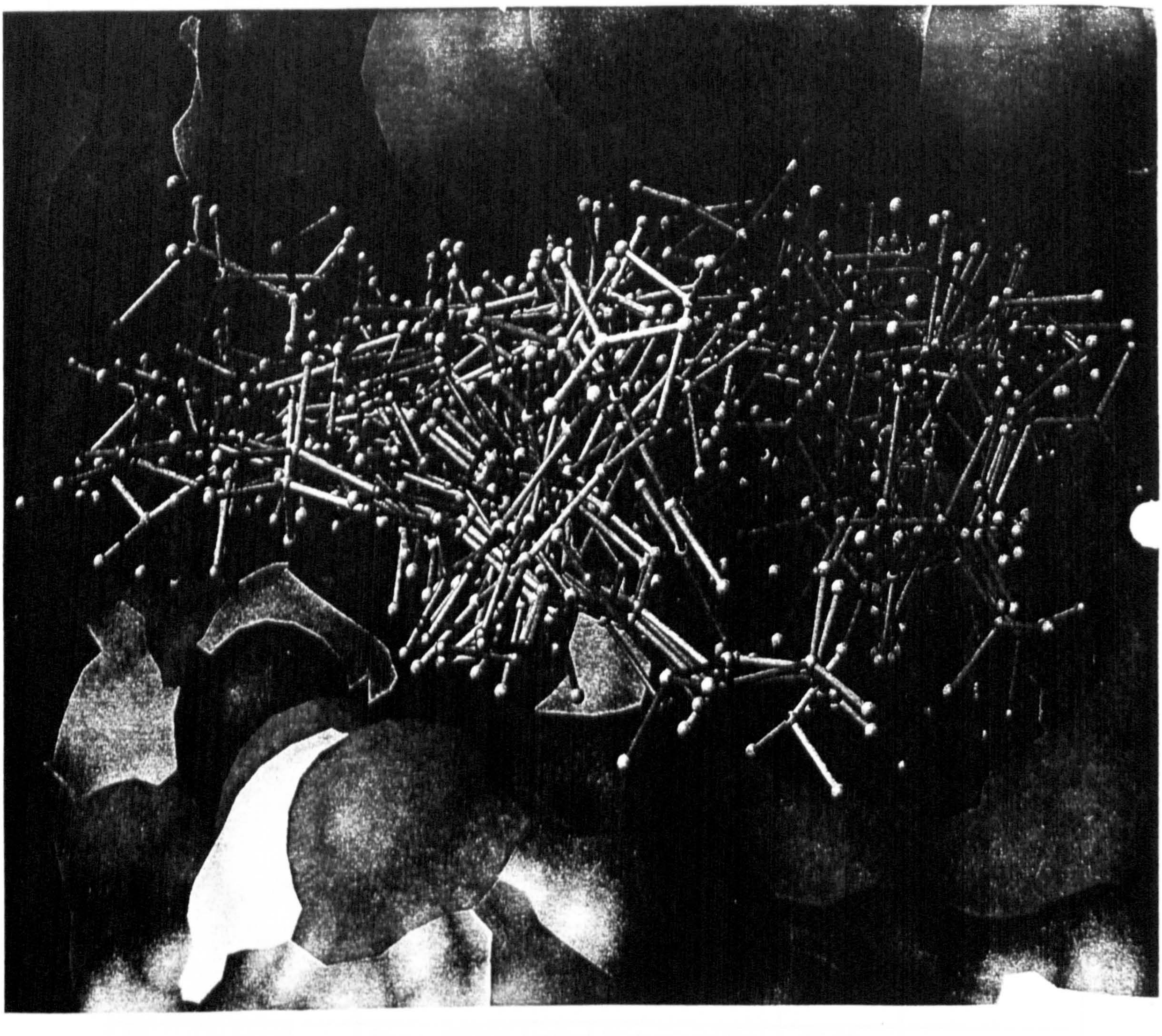


Fig. 5.2 A random sample of 100 ligands docked by the screening procedure in the active site of FGFR-TK.

Fig. 5.3 (a) The gel electrophoresis results of experimental tests for inhibition of kinase activity of FGFR-TK by 49 compounds selected in virtual database screen. Compounds 1 through 25 according to the index in Table 5.2

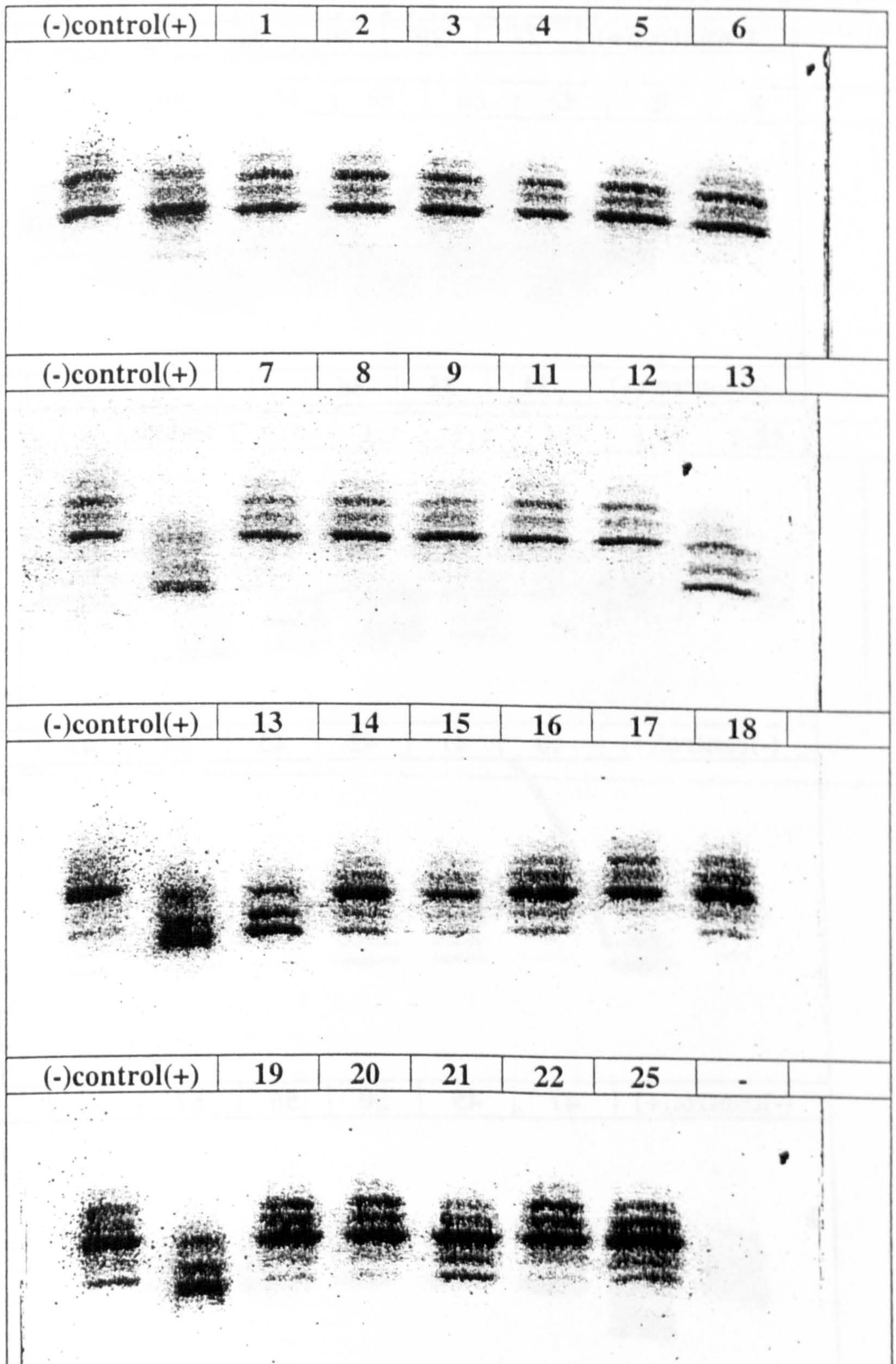


Fig. 5.3 (b) The gel electrophoresis results of experimental tests for inhibition of kinase activity of FGFR-TK by 49 compounds selected in virtual database screen. Compounds 27 through 50,56 and 57 according to the index in Table 5.2

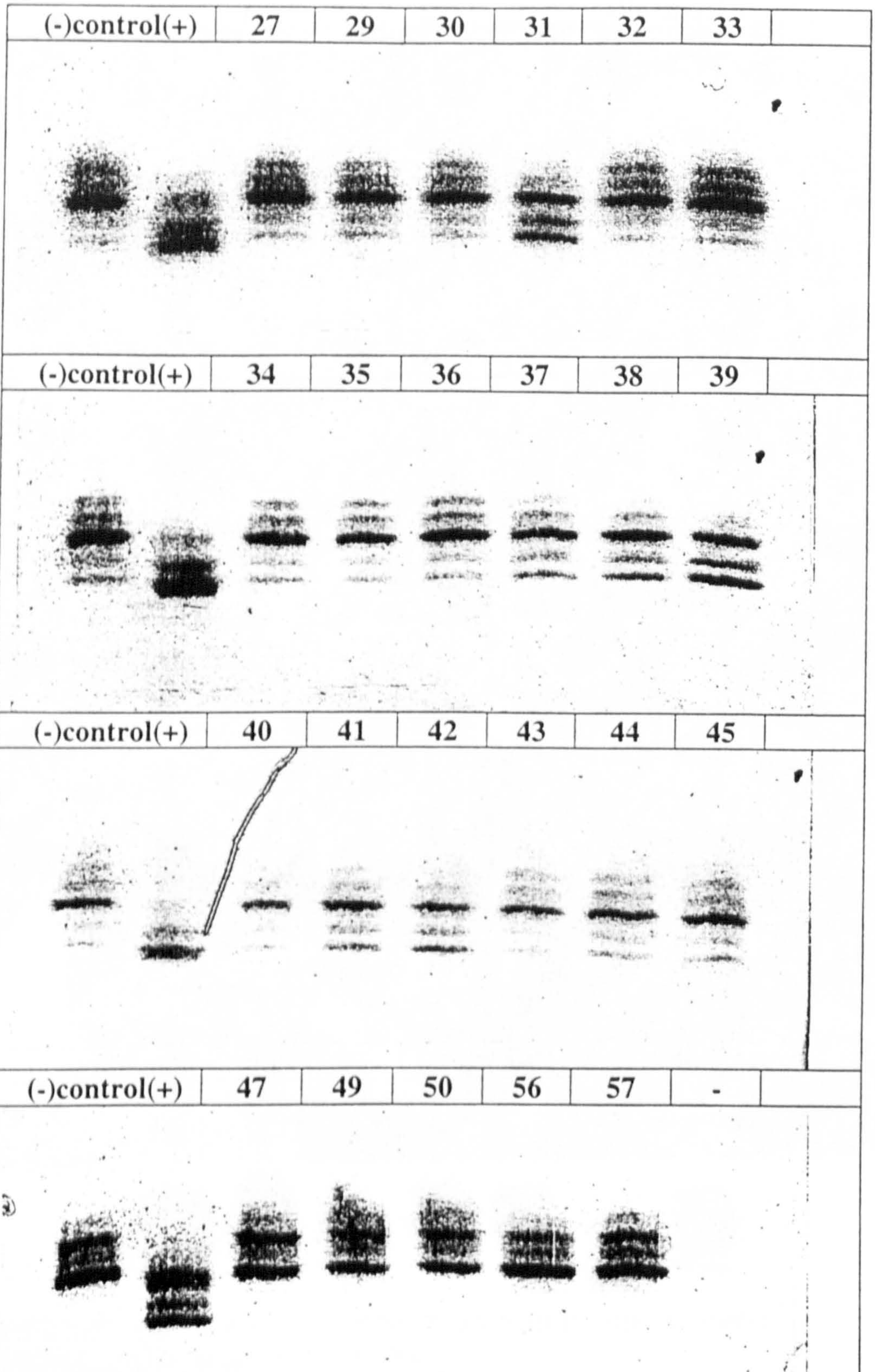
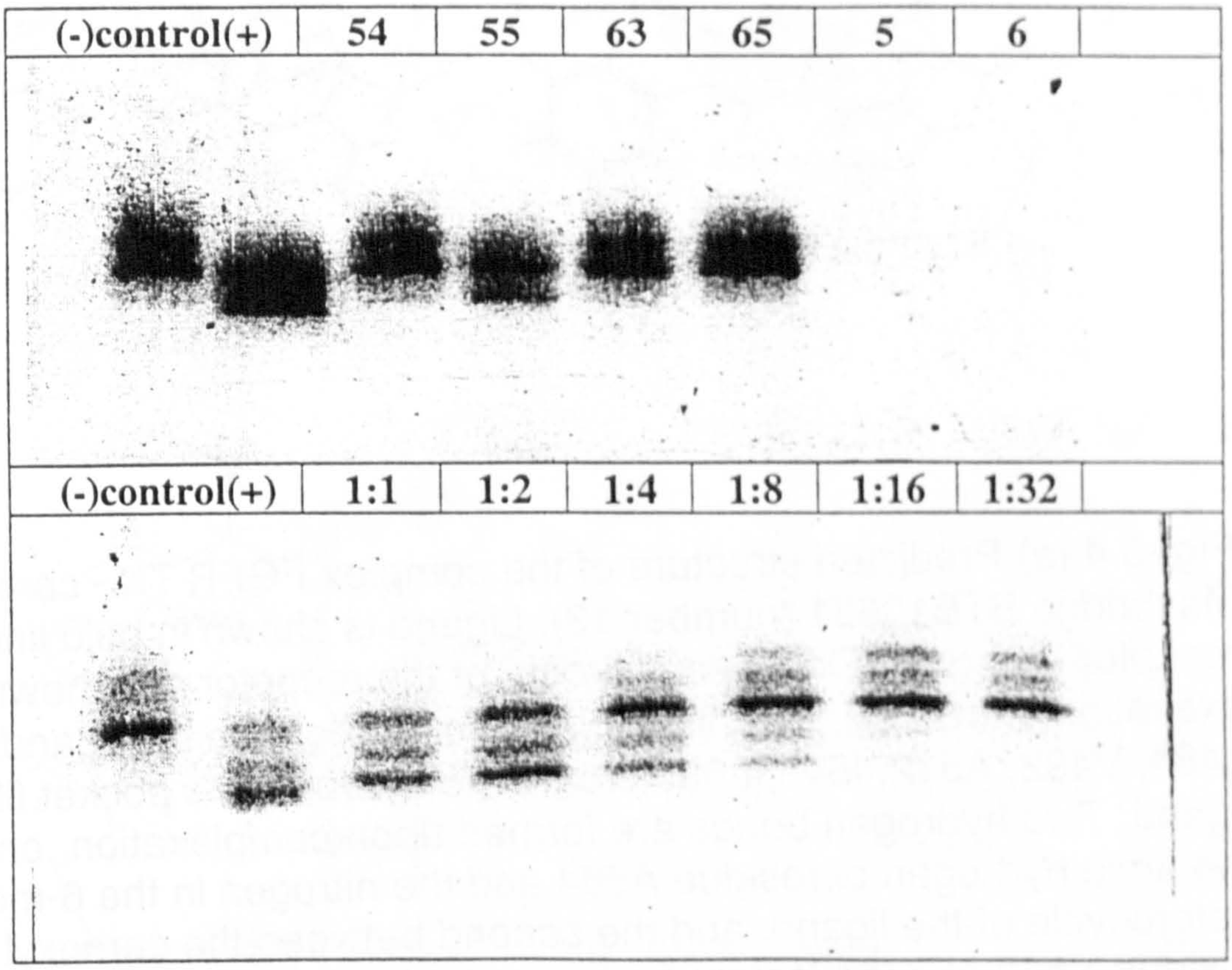


Fig. 5.3 (c) The gel electrophoresis results of experimental tests for inhibition of kinase activity of FGFR-TK by 49 compounds selected in virtual database screen. Compounds 54,55,63,65 according to the index in Table 5.2. Also shown are the results of dilution test of compound #13. The decline of the activity at lower inhibitor concentration can be clearly seen, with a median at 1:4 dilution.



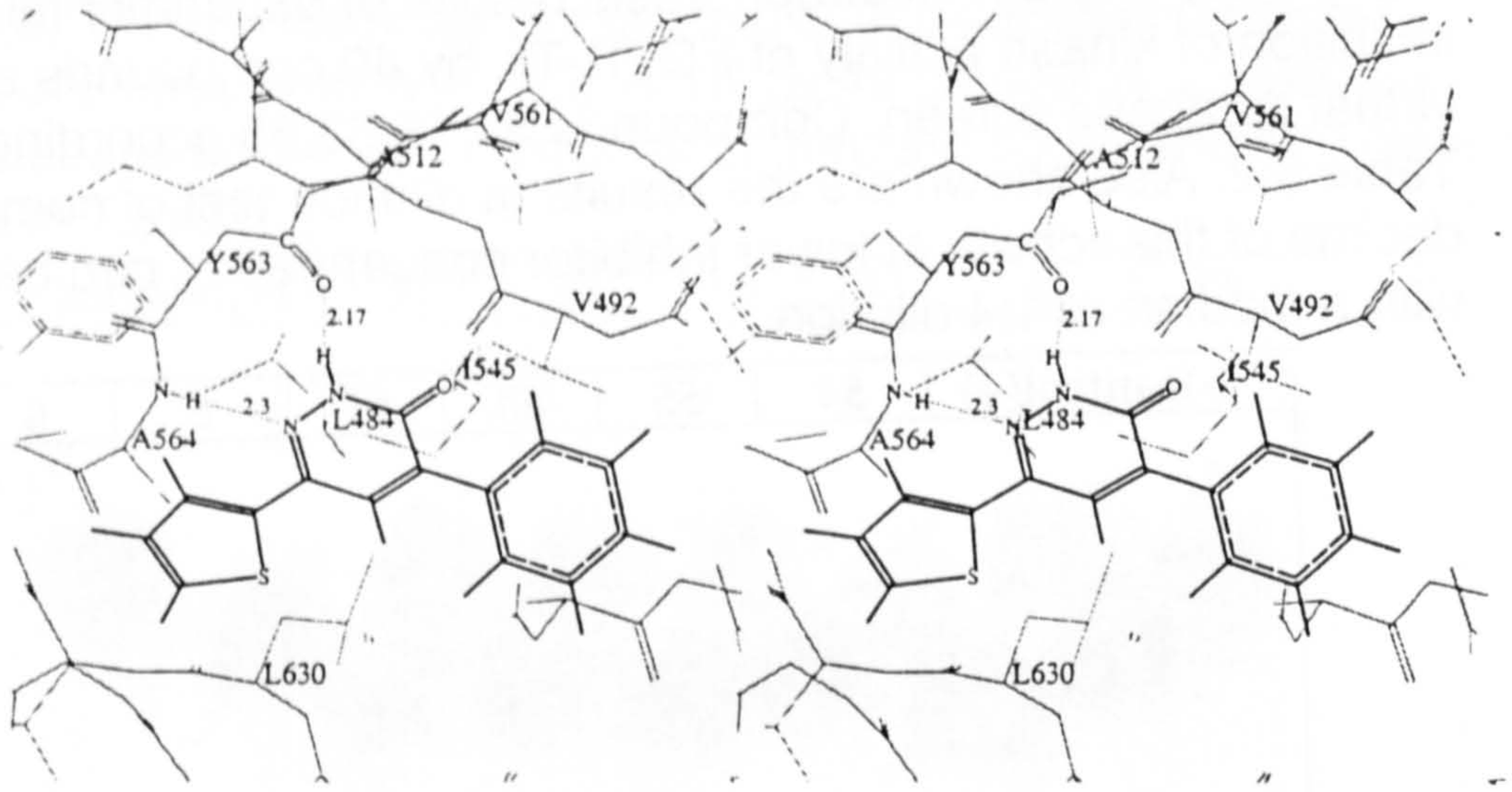


Fig. 5.4 (a) Predicted structure of the complex FGFR TK - compound Maybridge BTB10631 (number 13). Ligand is shown in bold lines, while the receptor is in gray. Only heavy atoms of the receptor are shown, with the exception of the hydrogen forming hydrogen bond to the ligand. Residues L484, V492, A512, I545, L630 create the hydrophobic pocket filled by the ligand. Two hydrogen bonds are formed upon complexation, one between the amid hydrogen of residue A564 and the nitrogen in the 6-membered heterocycle of the ligand, and the second between the carboxyl oxygen of residue Y563 and the lactam hydrogen in the same heterocycle. The stereo image pair is shown.

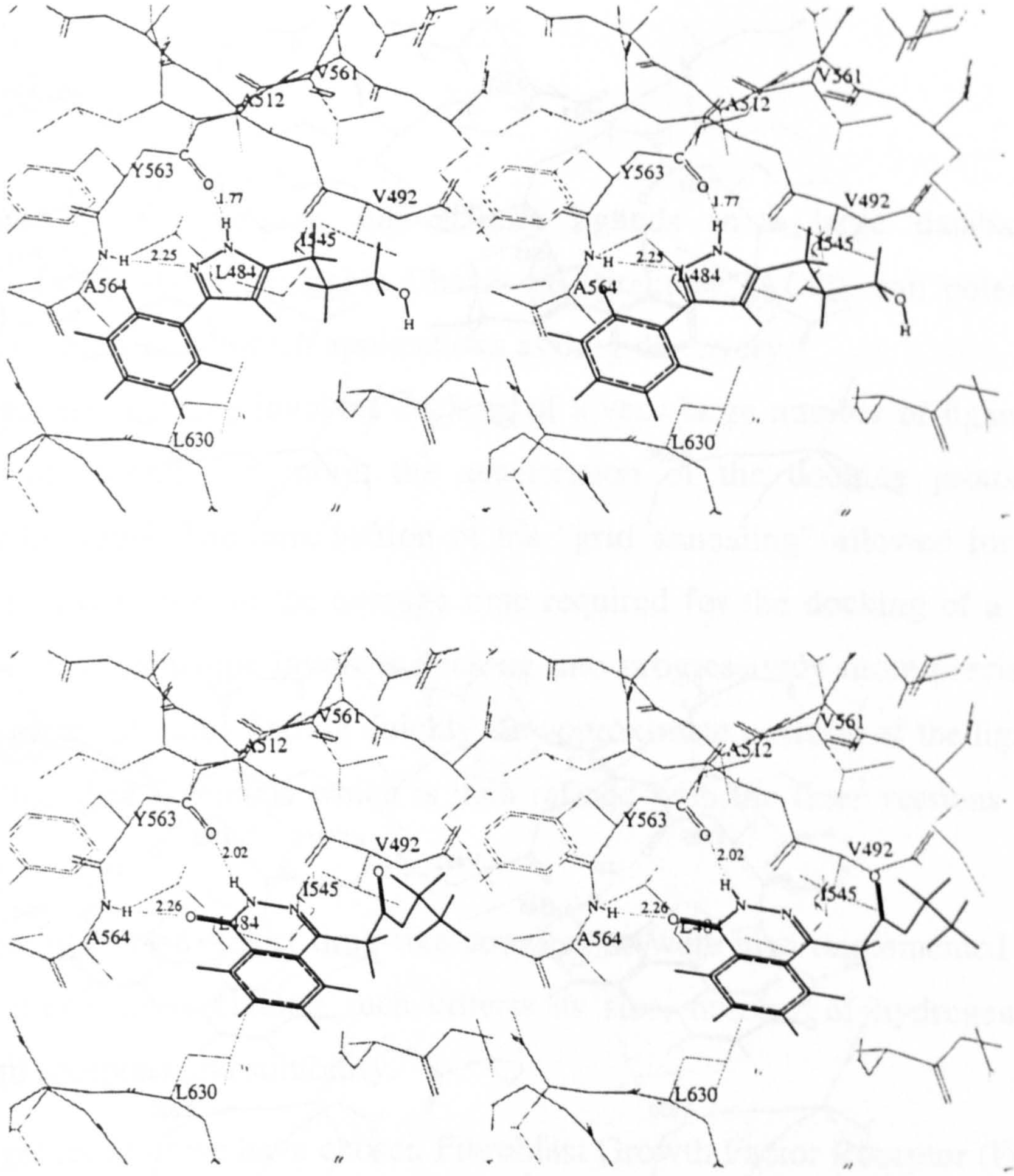


Fig 5.4 (b),(c) Predicted structure of the complexes FGFR TK - compound Maybridge CD09631 and KM06915 (number 21 and 31).

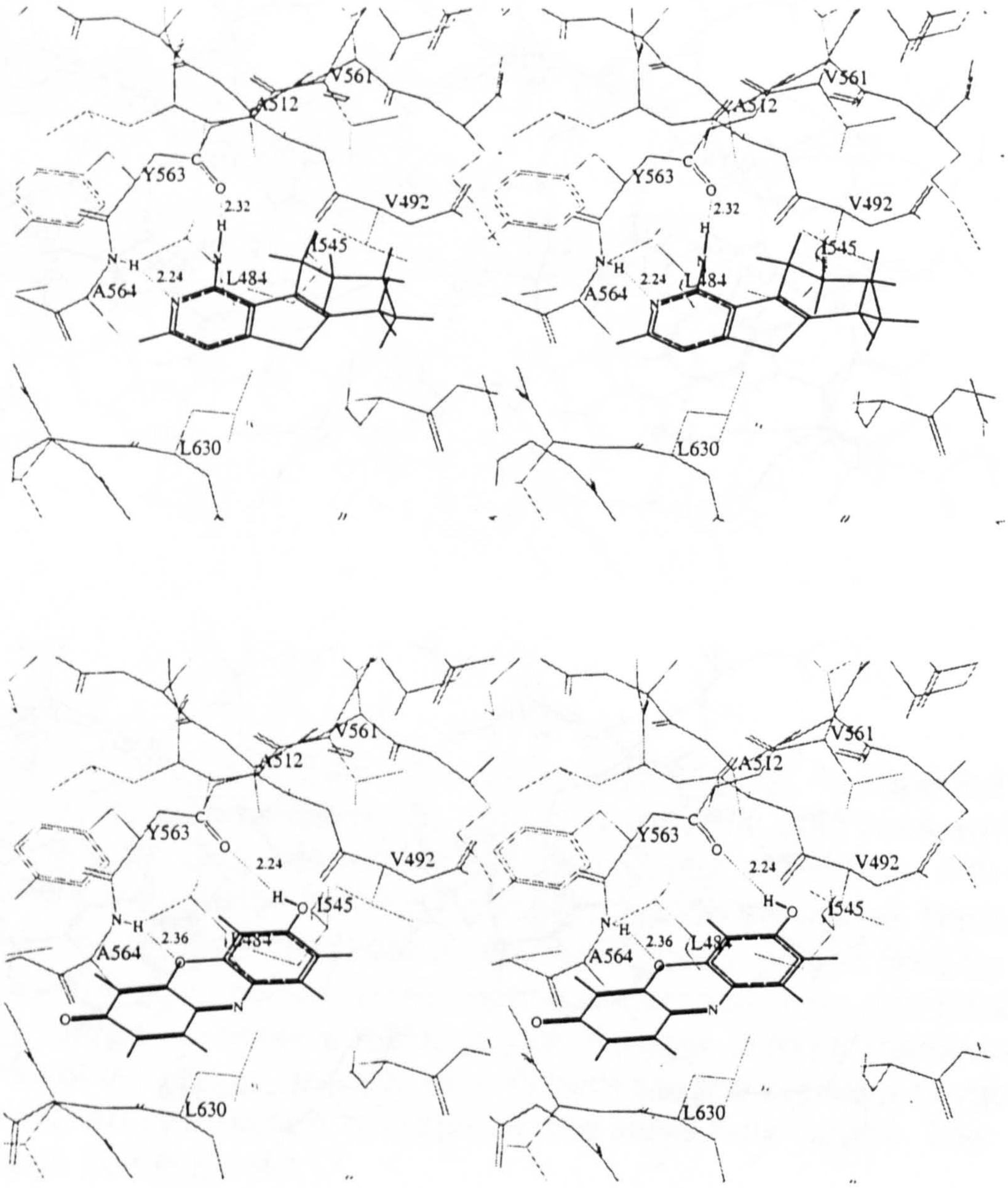


Fig 5.4 (d),(e) Predicted structure of the complexes FGFR TK - compound Maybridge SEW04081 and Sigma L7512 lacmoid (number 39 and 55).

5.4 Discussion

Automated identification of high-affinity ligands in a large database of compounds, such as the Available Chemicals Directory (ACD), can potentially become a valuable tool in such applications as drug discovery.

Since database scanning involves docking of a very large number of ligands, of the order of 100000 and more, the acceleration of the docking protocol is extremely desirable. The introduction of the "grid annealing" allowed for more than two-fold reduction in the average time required for the docking of a single compound. This technique involves docking into progressively more precise and rigid grid potentials, thus finding quickly the approximate position of the ligand in strongly smoothed potentials which is then refined with the finer versions of the potentials.

A number of rules to select drug-like compounds were also implemented in the scanning procedure, including such criteria as size, number of hydrogen bond donors and acceptors and solubility.

As a target receptor we have chosen Fibroblast Growth Factor Receptor (FGF-R) Tyrosine Kinase (TK). This receptor plays an important role in a number of cancers and other diseases, and the collaboration with the group of Steve Hubbard in the Skirball Institute provided us an opportunity to test the predictions experimentally. Dr. Hubbard also kindly provided the X-ray structure of FGF-R TK.

The scanning of 150,000 compounds identified ~4000 ligands with a favourable binding score and other features. From the complexes of several known inhibitors of FGF-R TK we could deduce the importance of the formation of two specific hydrogen bonds for the strong binding. This criteria was used as a further filter, which yielded ~150 ligands. We were pleasantly surprised to find out that about

10% of them were known inhibitors of TK or closely related compounds. Successful identification of these ligands confirmed that our scanning procedure indeed distinguishes ligands from the very large number of other compounds. 5 out of 49 experimentally tested novel ligand candidates showed inhibition activity. Predicted bound conformations for these ligands showed a number of favourable interactions.

Still, the number of false positives was quite high, reflecting the imperfections of the binding discrimination potential. Ways to the further improvement in discrimination can be suggested by in-depth analysis of the compounds erroneously predicted as ligands.

Lead discovery is the first step in the development of potent inhibitors. Improvement of the binding affinity of the discovered ligands can be achieved through chemical modifications, and docking simulations for the ligand derivatives can be used to suggest possible enhancements.

6. Conclusions

6.1 Overview

The research project described in this thesis involved three main stages covered in Chapters 3,4 and 5. The first stage was primarily focused on the full atom docking predictions for individual ligands. The use of internal co-ordinate mechanics made feasible the simulations involving flexibility of the receptor side-chains as well as the flexible ligand and achieved thorough sampling of the conformational space of the system. A convenient measure of the solution accuracy was developed.

The second stage involved the implementation and improvement of the grid potential docking methodology as well as development of the discrimination potential which was capable of identifying of the ligands binding to a particular receptor in a large pool of compounds. Grid representation of the receptor was essential to make the docking procedure fast enough for docking of a large number of putative ligands. Such representation also facilitated modifications of the potentials which further accelerate the convergence of the docking simulations and allow for more accurate solutions.

In third stage the techniques developed were further improved and applied to the real case of inhibitor discovery for fibroblast growth factor receptor tyrosine kinase. FGFR-TK is an important target for anti-cancer drugs.

6.2 Summary of Results

6.2.1 Flexible docking of individual ligands in full-atom representation

Using global minimisation of the free energy of the complex in the internal co-ordinate space eight protein-ligand complexes were simulated with flexible

ligand and receptor side-chains. Monte-Carlo minimisation procedure used two types of random moves, a pseudo Brownian positional move and a Biased-Probability multi-torsion move, each accompanied by full local energy minimisation. The best docking solutions were further ranked according to the interaction energy which included intramolecular deformation energies of both receptor and ligand, the interaction energy, surface tension, side-chain entropic contribution and an electrostatic term evaluated as a boundary element solution of the Poisson equation with the molecular surface as a dielectric boundary. The geometrical accuracy of the docking solutions ranged from 30% to 70% according to the relative displacement error measure at a 1.5Å scale.

6.2.2 Grid docking and ligand discrimination

A fast and flexible docking protocol utilising the grid potential representation of the receptor molecule was developed. A sophisticated binding discrimination potential was implemented, which included all major contributions to the binding energy, such as electrostatic energy, hydrophobicity, hydrogen bonding, softened Van der Waals and entropic contributions. The docking protocol was tested on a set of 51 known structures of complexes for 23 receptor molecules. 35 predicted structures had a correct overall binding mode with RMSD of 3Å or less from the native structure and 26 structures were closer than 2Å with most of the details of binding conserved. Exhaustive cross-docking of 23 receptors and 63 ligand compounds produced putative complex structures for all ligand-receptor combinations. Generated complex structures were used to optimise the discrimination potential for identification of the native pairs. Optimised potential successively identified native ligands for 13 receptors and in all but two cases at least one native ligand was within the first 3 selected compounds.

6.2.3 Database screening and discovery of novel inhibitors of FGFR-TK

Large database of the commercially available compounds containing over 150000 was screened using the discrimination potential optimised for selectivity on a diverse set of protein-ligand complexes from PDB. Docking protocol produced putative complex structures of all database compounds and FGFR-TK domain, an important target in anti-cancer drug design. 49 putative ligands picked by the screening protocol were purchased and experimentally tested. Five of them showed detectable activity and four had sufficient affinity to compete with natural ligand ATP. At least two ligands belong to entirely novel families of tyrosine kinase inhibitors.

6.3 Future work

The results obtained in this work suggest promising further research in several directions. One important development would be the addition of the lead optimisation protocol which would attempt to evaluate possible chemical modifications of the discovered ligands with a view to improving the binding. Such a protocol would involve a combinatorial search of the derivatives of the lead using a library of chemical groups which can be attached to the lead at a number of positions in various combinations. All the derivatives constructed can be docked and evaluated by the procedures already developed. Such a procedure can then be used to improve the binding affinity of the FGFR-TK inhibitors that we have discovered.

As the screening protocol is currently rather time-consuming, further acceleration of the docking routine is also desirable.

Further improvement of the docking and discrimination potentials is another important direction, as the selectivity is still rather low – only 1 in 10 compounds selected worked as inhibitors. One possibility for the improvement of the potential

is the better treatment of hydrogen bonds. In its current form the directionality of hydrogen bonding is rather poor, especially in the case of the acceptor atoms of the ligand, where the geometry of the lone electron pairs is completely ignored. This problem can be circumvented by the introduction of the additional centres attached to the acceptor atoms and centred along the directions of the lone pairs. The hydrophobic potential currently treats all atoms only as totally hydrophobic or hydrophilic. Introduction of intermediate states may result in better treatment of such groups as aromatic rings, which are less hydrophobic than aliphatic groups. These enhancements should help the scanning procedure to identify reliably the most potent ligands and, ultimately, novel drugs.

7. Publications

Totrov, M., and Abagyan, R., (1999). Derivation of sensitive discrimination potential for virtual ligand screening *Proceedings of RECOMB99.* (in press)

Abagyan, R., and Totrov, M. (1999). Ab Initio Folding of Peptides by the Optimal-Bias Monte Carlo Minimization Procedure *J. Comp. Phys.*, **151**, (in press).

Totrov, M. & Abagyan, R., (1997). Flexible protien-ligand docking by global energy optimization in internal coordinates Proteins, Suppl. **1**, 215-220.

Abagyan, R., Batalov, S., Cardozo, T., Totrov, M., Webber, J. and Zhou, Y. (1997). Homology Modeling with Internal Coordinate Mechanics: Deformation Zone Mapping and Improvements of Models via Conformational Search. Proteins: Struct. Funct. & Gen., Suppl. **1**, 29-37.

Abagyan, R. & Totrov, M. (1997). Contact Area Difference (CAD): A robust measure to evaluate accuracy of protein models. J. Mol. Biol., **268**, 678-685.

Totrov, M. & Abagyan, R.A. (1996). The contour-buildup algorithm to calculate the analytical molecular surface. J. Struct. Biol., **116**, 138-143.

N.C.J. Strynadka, M. Eisenstein, E. Katchalski-Katzir, B.K.Shoichet, I.D. Kuntz, R. Abagyan, M. Totrov, J. Janin, J. Cherfils, F. Zimmerman, A. Olson, B. Duncan, M. Rao, R. Jackson, M. Sternberg, and M.N.G. James (1996). Molecular docking programs successfully predict the binding of a beta-lactamase inhibitory protein to TEM-1 beta-lactamase. Nature Struct. Biol., **3**, 233-239.

Chalikian,T.V., Totrov, M.M., Abagyan, R.A., Breslauer, K.J. (1996). The hydration of globular proteins as derived from volume and compressibility measurements: cross correlating thermodynamic and structural data. J. Mol. Biol., **260**, 588-603.

8. References

1. Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). The protein data bank: a computer-based archival file for macromolecular structures. J. Mol. Biol., **112**, 535-542.

2. Beddell C.R., Goodford P.J., Norrington F.E., Wilkinson S., Wootton R. (1976). Compounds designed to fit a site of known structure in human haemoglobin. Br J Pharmacol, **57(2)**, 201-209.

3. Kuntz, I.D., Blaney, J.M., Oatley, S.J., Langridge, R. & Ferrin, T.E. (1982). A geometric approach to macromolecule-ligand interactions. J. Mol. Biol., **161**, 269-288.

4. DesJarlais R.L., Sheridan R.P., Dixon J.S., Kuntz I.D., Venkataraghavan R. (1986). Docking flexible ligands to macromolecular receptors by molecular shape J Med Chem, **29**, 2149-2153.

5. Bacon, D.J. & Moult, J. (1992). Docking by Least-squares Fitting of Molecular Surface Patterns. J. Mol. Biol., 225, 849-858.

6. Leach, A.R. & Kuntz, I.D. (1992). Conformational analysis of flexible ligands in macromolecular receptor sites. J. Comp. Chem., **13**, 733-748.

7. Katchalski-Katzir, E., Shariv, I., Eisenstein, M., Friesem, A.A., Aflalo, C., Vakser, I.A (1992). Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques. Proc Natl Acad Sci U S A, **89**, 2195-9.

8. Helmer-Citterich M, Tramontano A (1994). PUZZLE: a new method for automated protein docking based on surface shape complementarity. ???, **235(3)**, 1021-1031.

9. Fischer, D., Lin, S.L., Wolfson, H.L., Nussinov, R (1995). A geometry-based suite of molecular docking processes. J. Mol. Biol., 248, 459-477.

10. Jiang, F. & Kim S.-H. (1991). Soft docking J. Mol. Biol., **219**, 79-102.

11. Walls, P.H. & Sternberg, M.J.E. (1992). New algorithm to model protein-protein recognition based on surface complementarity. J. Mol. Biol., **228**, 277-297.

12. DiNola, A., Raccatano, D. & Berendsen, H. (1994). Molecular dynamics simulation of the docking of substrates to proteins. Proteins, **19**, 174-182.

13. Abagyan, R.A., Totrov, M.M. & Kuznetsov, D.A. (1994). ICM: a new method for structure modeling and design: Applications to docking and structure prediction from the distorted native conformation. J. Comp. Chem., **15**, 488-506.

14. Nilges, M. & Brunger, A. (1993). Successful Prediction of the Coiled Coil Geometry of the GCN4 Leucine Zipper Domain by Simulated Annealing: Comparison to the X-Ray Structure. Proteins, **15**, 133-146.

15. Shoichet, B.K. & Kuntz, I.D. (1991). Protein Docking and Complementarity. J. Mol. Biol., **221**, 327-346.

16. Bohm, H.J. (1994). On the use of LUDI to search the Fine Chemicals Directory for ligands of proteins of known three-dimensional structure. J Comput Aided Mol Des, 8(5), 623-632.

17. Bohm, H.J. (1994). The development of a simple empirical scoring function to estimate the binding constant for a protein-ligand complex of known three-dimensional structure. J Comput Aided Mol Des, 8(3), 243-256.

18. Jain AN (1996). Scoring noncovalent protein-ligand interactions: a continuous differentiable function tuned to compute binding affinities. J Comput Aided Mol Des, 10(5), 427-440.

19. Totrov, M.M. & Abagyan, R.A. (1994). Detailed ab initio prediction of lysozyme-antibody complex with 1.6A accuracy. Nature Structural Biology, **1**, 259-263.

20. Davis, M.E. & McCammon, J.A. (1990). Electrostatics in Biomolecular Structure and Dynamics. Chem. Rev., **90**, 509-521.

21. Warshel, A. & Russell, S.T. (1984). Calculation of electrostatic interactions in biological systems and in solution. Quart.Rev.Biophys., **17**, 283-422.

22. Tanford, C. & Kirkwood, J.G. (1957). Theory of Protein Titration Curves. I. General Equations for Impenetrable Spheres. J. Amer. Chem. Soc., **79**, 5333-5339.

23. Harvey, S. (1989). Treatment of Electrostatic Effects in Macromolecular Modeling Proteins, **5**, 78-92.

24. Nicholls, A. & Honig, B. (1991). A Rapid Finite Difference Algorithm, Utilizing Successive over-Relaxation to Solve the Poisson-Boltzmann Equation. J. Comput. Chem., **12**, 435-445.

25. Kauzmann, W. (1959). Some factors in the interpretation of protein denaturation. Adv. Protein Chem., **14**, 1-63.

26. Chothia, C. (1976). The nature of the accessible and buried surfaces in proteins. J. Mol. Biol., **105**, 1-12.

27. Sharp, K.A., Nicholls, A., Fine, R.F. & Honig, B. (1991). Reconciling the Magnitude of the Microscopic and Macroscopic Hydrophobic Effects. Science, **252**, 106-109.

28. Sitkoff, D., Sharp, K.A., and Honig B. (1994). Accurate Calculation of Hydration Free Energies Using Macroscopic Solvent Models. Journal of Physical Chemistry, 98(7), 1978-1988.

29. Horton N, Lewis M (1992). Calculation of the free energy of association for protein complexes. Protein Sci., 1(1), 169-181.

30. Brooks, B.R., Bruccoleri, R.E., Olafson, B.D., States, D.J. Swaminathan, S. & Karplus, M. (1983). CHARMM: A Program for Macromolecular Energy, Minimization and Dynamics Calculations. J. Comput. Chem., **4**, 187-217.

31. Weiner, S.J., Kollman, P.A., Case, D.A., Chandra Singth, U., Ghio, C., Alagona, G., Prefeta, S., Jr. & Wiener, P. (1984). A New Force Field for Molecular Mechanical Simulation of Nucleic Acids and Proteins. J. Amer. Chem. Soc., **106**, 765-783.

32. Halgren, T.A. (1995). Merck Molecular Force Field. I.-V. J. Comp. Chem., **17**, 490-641.

33. Momany, F.A., McGuire, R.F., Burgess, A.W. & Scheraga, H.A. (1975). Energy Parameters in Polypeptides. VII. Geometric Parameters, Partial Atomic Charges, Nonbonded Interactions, Hydrogen Bond Interactions, and Intrinsic Torsional Potentials for the Naturally Occurring Amino Acids. J. Phys. Chem., **79**, 2361-2381.

34. Connolly, M.L. (1986). Shape Complementarity at the Hemoglobin " 1° 1 Subunit Interface. Biopolymers, **25**, 1229-1247.

35. Nemethy, G., Gibson, K.D., Palmer, K.A., Yoon, C.N., Paterlini, G., Zagari, A., Rumsey, S. & Scheraga, H.A. (1992). Energy Parameters in Polypeptides. 10. Improved geometric parameters and nonbonded interactions for use in the ECEPP/3 algorithm, with application to proline-containing peptides. J. Phys. Chem., **96**, 6472-6484.

36. Ippolito JA, Alexander RS, Christianson DW (1990). Hydrogen bond stereochemistry in protein structure and function. J Mol Biol , **215(3)**, 457-471.

37. Goodford, P.J. (1985). A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. J. Med. Chem., **28**, 849-875.

38. Miller MD, Kearsley SK, Underwood DJ, Sheridan RP (1994). FLOG: a system to select 'quasi-flexible' ligands complementary to a receptor of known three-dimensional structure. J Comput Aided Mol Des, **8(2)**, 153-174.

39. Yang, A.-S., Sharp, K.A. & Honig, B. (1992). Analysis of the Heat Capacity Dependence of Protein Folding. J. Mol. Biol., **227**, 889-900.

40. Abagyan, R.A. & Totrov, M.M. (1994). Biased Probability Monte Carlo Conformational Searches and Electrostatic Calculations for Peptides and Proteins. J.Mol.Biol., **235**, 983-1002.

41. Cherfils, J., Duquerroy, S. & Janin, J. (1991). Protein-protein recognition analyzed by docking simulation. Proteins, **11**, 271-280.

42. Kearsley, S.K., Underwood, D.J., Sheridan, R.P., Miller, M.D. (1994). Flexibases: a way to enhance the use of molecular docking methods. J Comput Aided Mol. Design, 8, 565-82.

43. Welch, W., Ruppert, J., Jain, A.J. (1996). Hammerhead: fast, fully automated docking of flexible ligands to protein binding sites. Chemistry & Biology, **3**, 449-462.

44. Bohm, H.J. (1992). LUDI: rule-based automatic design of new substituents for enzyme inhibitor leads. J. Comput. Aided Mol. Design, 6, 593-606.

45. Metropolis, N.A., Rosenbluth, A.W., Rosenbluth, N.M., Teller, A.H., Teller, E. (1953). Equation of State calculations by Fast Computing Machines. J. Chem. Phys., **21**, 1087-1092.

46. Li, Z. & Scheraga, H.A. (1987). Monte Carlo-Minimization Approach to the Multiple-Minima Problem in Protein Folding. Proc. Natl. Acad. Sci. USA, 84, 6611-6615.

47. Abagyan, R.A. & Argos, P. (1992). Optimal protocol and trajectory visualization for conformational searches of peptides and proteins. J. Mol. Biol., 225, 519-532.

48. G. Jones, P. Willett, R.C. Glen, A.R. Leach, and R. Taylor (1997). Development and validation of a genetic algorithm for flexible docking. J. Mol. Biol., 267, 727-748.

49. Rossky, P.J., Doll, J.D. & Friedman, H.L. (1978). Brownian Dynamics as Smart Monte Carlo Simulation. J. Chem. Phys., 69, 4628-4633.

50. Kozack RE, Subramaniam S (1993). Brownian dynamics simulations of molecular recognition in an antibody-antigen system. Protein Sci., 2(6), 915-926.

51. Krystek S, Stouch T, Novotny J (1993). Affinity and specificity of serine endopeptidase-protein inhibitor interactions. Empirical free energy calculations based on X-ray crystallographic structures. J Mol Biol, 234(3), 661-679.

52. Vajda S, Weng Z, Rosenfeld R, DeLisi C (1994). Effect of conformational flexibility and solvation on receptor-ligand binding free energies. Biochemistry, 33(47), 13977-13988.

53. Lee FS, Chu ZT, Bolger MB, Warshel A (1992). Calculations of antibody-antigen interactions: microscopic and semi-microscopic evaluation of the free energies of binding of phosphorylcholine analogs to McPC603. Protein Eng., 5(3), 215-228.

54. Rastelli G., Thomas B., Kollman P.A., Santi D.V. (1995). Insight into the specificity of thymidilate synthase from molecular dynamics and free energy perturbation calculations J Am Chem Soc, 117, 7213-7227.

55. Andrews PR, Craik DJ, Martin JL (1984). Functional group contributions to drug-receptor interactions. J Med Chem, 27(12), 1648-1657.

56. Bohm HJ (1998). Prediction of binding constants of protein ligands: a fast method for the prioritization of hits obtained from de novo design or 3D database search programs. J Comput Aided Mol Des, 12(4), 309-323.

57. Lee, B., Richards, F.M. (1971). The interpretation of protein structures: Estimation of static accessibility. J. Mol. Biol., 55, 379-400.

58. Shrake, A. & Rupley, J.A. (1973). Environment and Exposure to Solvent of Protein Atoms. Lysozyme and Insulin. J. Mol. Biol., 79, 351-371.

59. Richards, F.M. (1977). Areas, volumes, packing and protein structure. Ann. Rev. Biophys. Bioeng., 6, 151-176.

60. Lorensen, W., Cline, H. (1987). Marching cubes: A high resolution 3D surface construction algorithm. Computer Graphics, 21, 163-169.

61. Connolly, M.L. (1983). Analytical molecular surface calculation. J. Appl. Cryst., 16, 548-558.

62. Totrov, M.M. & Abagyan, R.A. (1996). The contour-buildup algorithm to calculate the analytical molecular surface. J. Struct. Biol., 116, 138-143.

63. Juffer, A.H., Botta, E.F.F., van Keulen, B.A.M., van der Ploeg, A. & Berendsen, H.J.C. (1991). The electric potential of a macromolecule in a solvent: a fundamental approach. J. Comput. Phys., **97**, 144-171.

64. Zauhar, R.J. and Morgan, R.S. (1985). A new method for computing the macromolecular electric potential. J.Mol.Biol., **186**, 815-820.

65. Purisima, E.O. &, Nilar, S.H. (1995). A Simple Yet Accurate Boundary Element Method for Continuum Dielectric Calculations J. Comp. Chem., **16**, 864-870?.

66. William H.Press at al. (1992). Numerical recipes in C: the art of scientific computing. Cambridge University Press, , .

67. Abagyan, R. & Totrov, M. (1997). Contact Area Difference (CAD): A robust measure to evaluate accuracy of protein models. J. Mol. Biol., **268**, 678-685.

68. Gulukota, K., Vajda, S. & Delisi, C. (1996). Peptide Docking Using Dynamic Programming. J. Comp. Chem., 17, 418-428.

69. Rarey, M., Kramer, B., Lengauer, T., Klebe, G (1996). A fast flexible docking method using an incremental construction algorithm. J Mol Biol, **261**, 470-489.

70. Leach, A.R. (1994). Ligand docking to proteins with discrete side-chain flexibility. J. Mol. Biol., **235**, 345-356.

71. Rosenfeld, R., Vajda, S. & DeLisi, C. (1995). Flexible docking and design. Annu. Rev. Biophys. Biomol. Struct., **24**, 677-700.

72. Miranker, A. & Karplus, M. (1991). Functionality maps of binding cites: a multiple copy simultaneous search method. Proteins: Struct., Funct. & Dyn., **11**, 29-34.

73. Luty, B.A., Wasserman, Z.R., Stouten, P.F.W., Hodge, C.N., Zacharias, M., McCammon, J.A. (1995). A molecular mechanics/grid method for evaluation of ligand-receptor interactions. J. Comput. Chem., **16**, 454-464.

74. Hartl, F.U. & Martin, J. (1992). Protein folding in the cell: the role of molecular chaperones Hsp70 and Hsp80. Ann. Rev. Biophys. Biomol. Struct., **21**, 293-322.

75. Goodsell A.S.; Olson A.J. (1990). Automated docking of substrates to proteins by simulated annealing. Proteins: Struct., Funct. \& Gen., **8**, 195-202.

76. Caflisch, A., Niederer, P., Anliker, M. (1992). Monte Carlo docking of oligopeptides to proteins. Proteins: Struct., Funct. \& Gen., **13**, 223-230.

77. Janin, J. (1995). Protein-protein recognition. Prog. Biophys. molec. Biol., **64**, 145-166.

78. N.C.J. Strynadka, M. Eisenstein, E. Katchalski-Katzir, B.K.Shoichet, I.D. Kuntz, R. Abagyan, M. Totrov, J. Janin, J. Cherfils, F. Zimmerman, A. Olson, B. Duncan, M. Rao, R. Jackson, M. Sternberg, and M.N.G. James (1996). Molecular docking programs successfully predict the binding of a beta-lactamase inhibitory protein to TEM-1 beta-lactamase. Nature Struct. Biol., **3**, 233-239.

79. Chothia, C., Lesk, A.M., Tramontano, A., Levitt, M., Smith-Gill, S.J., Air, G., Sheriff, S., Padlan, E.A., Davies, D., Tulip, W.R. (1989). Conformations of immunoglobulin hypervariable regions Nature, **342**, 877-883.

80. Allen, F.H., Kennard, O. (1993). 3D Search and Research Using the Cambridge Structural Database. Chemical Design Automation News, 8 (1), 1 & 31-37.

81. Hart, T.N. & Read, R.J. (1992). A multiple-start Monte Carlo docking model. Proteins: Struct., Funct. \& Gen., 13, 206-222.

82. Novotny, J., Bruccoleri, R.E., Saul, F.A. (1989). On the attribution of binding-energy in antigen-antibody complexes Mcpc-603. Biochemistry, 28, 4735-4749.

83. Honig, B., Nicholls, A. (1995). Classical Electrostatics in Biology and Chemistry. Science, 268, 1144-1149.

84. Zacharias, M., Luty, B.A., Davis, M.E., McCammon, J.A (1994). Combined conformational search and finite-difference Poisson-Boltzmann approach for flexible docking. Application to an operator mutation in the lambda repressor-operator complex. J. Mol. Biol., 238, 455-465.

85. Jackson, R.M., Sternberg, M.J.E. (1995). A continuum Model for Protein-Protein Interactions: Application to the Docking Problem. Journal of Molecular Biology, 250, 258-275.

86. Bharadwaj, A., Windemuth, A., Sridharan, S., Honig, B., Nicholls, A. (1995). The Fast Multipole Boundary Element Method for Molecular Electrostatics: An Optimal Approach for Large System. J. Comp. Chem., 16, 898-910.

87. Bamborough, P. & Cohen, F.E. (1996). Modeling protein-ligand complexes. Curr. Opin. Struct. Biol., 6, 236-241.

88. Lengauer, T. & Rarey, M. (1996). Computational methods for biomolecular docking. Curr. Opinion Struct. Biol., 6, 402-406.

89. Shortle, D. (1992). Mutational studies of protein structures and their stabilities. Q. Rev. Biophys., 25, 205-250.

90. Marrone TJ, Briggs JM, McCammon JA (1997). Structure-based drug design: computational advances. Annu Rev Pharmacol Toxicol, 37, 71-90.

91. Totrov, M. & Abagyan, R., (1997). Flexible protien-ligand docking by global energy optimization in internal coordinates. Proteins, Suppl. 1, 215-220.

92. Shoichet, B.K., Stroud, R.M., Santi, D.V., Kuntz, I.D., Perry, K.M. (1993). Structure-based discovery of inhibitors of thymidylate synthase. Science, 259, 1445-1450.

93. Shoichet, B.K., Bodian, D.L. & Kuntz, I.D. (1992). Molecular docking using shape descriptors. J. Comp. Chem., 13, 380-397.

94. Hoffman LR, Kuntz ID, White JMStructure-based identification of an inducer of the low-pH conformationalchange in the influenza virus hemagglutinin: irreversible inhibition ofinfectivity. (1997). J. Virol, 71(11), 8808-8820.

95. Basilico C, Moscatelli D. (1992). The FGF family of growth factors and oncogenes. Adv. Cancer Res., 59, 115-165.

96. Naski MC, Wang Q, Xu J, Ornitz DM (1996). Graded activation of fibroblast growth factor receptor 3 by mutations causing achondroplasia and thanatophoric dysplasia. Nature Genetics, 13(2), 233-237.

97. Leung HY, Gullick WJ, Lemoine NR (1994). Expression and functional activity of fibroblast growth factors and their receptors in human pancreatic cancer. Int J Cancer, **59(5)**, 667-675.

98. Mohammadi et al., (1997). Structures of the Tyrosine Kinase Domain of Fibroblast Growth Factor Receptor in Complex with Inhibitors. Science, **276** , RP

99. Collett MS, Erikson RL (1978). Protein kinase activity associated with the avian sarcoma virus src gene product. Proc Natl Acad Sci, **75(4)**, 2021-2024.

100. Eckhart W, Hutchinson MA, Hunter T (1979). An activity phosphorylating tyrosine in polyoma T antigen immunoprecipitates. Cell, **18(4)**, 925-933.

101. Myers MR, He W, Hulme C (1997). Inhibitors of tyrosine kinases involved in inflammation and autoimmune disease. Current Pharmaceutical Design, **3**, 473-502.

102. Fecik RA, Frank KE, Gentry EJ, Menon SR, Mitscher LA, Telikepalli H (1998). The search for orally active medications through combinatorial chemistry. Med Res Rev, **18(3)**, 149-185.

103. Simmerling, C.L., Elber, R. (1995). Computer determination of peptide conformations in water: different roads to structure. Proc Natl Acad Sci USA, **92**, 3190-3193.

104. McCammon, J.A., Wolynes, P.G. & Karplus, M. (1979). Picosecond Dynamics of Tyrosine Side Chains in Proteins. Biochemistry, **18**, 927-942.

105. Pickersgill, R.W. (1988). A rapid method of calculating charge-charge interaction energies in proteins. Prot. Eng., **2**, 247-248.

106. Friedman, H.L. (1975). Image approximation to the reaction field. Molecular Physics, **29**, 1533-1543.

107. Schaefer, M. & Froemmel, C. (1990). A Precise Analytical Method for Calculating the Electrostatic Energy of Macromolecules in Aqueous Solution. J. Mol. Biol., **216**, 1045-1066.

108. Noguti, T. & Go, N. (1985). Efficient Monte Carlo Method for Simulation of Fluctuating Conformations of Native Proteins. Biopolymers, **24**, 527-546.

109. Vanderbilt, D. & Louie, S.G. (1984). A Monte Carlo Simulated Annealing Approach to Optimization over Continuous Variables. J. Comp. Phys., **56**, 259-271.

110. Shin, J.K. & Jhon, M.S. (1991). High Directional Monte Carlo Procedure Coupled with the Temperature Heating and Annealing as a Method to Obtain the Global Energy Minimum Structure of Polypeptides and Proteins. Biopolymers, **31**, 177-185.

111. Clearwater, S.H., Huberman, B.A., Hogg, T. (1991). Cooperative solution of constraint satisfaction problems. Science, **254**, 1181-1183.

112. Unger, R. and Moult, J. (1993). Genetic algorithms for protein folding simulations. J. Mol. Biol., **231**, 75-81.

113. Dandekar, T. and Argos, P. (1994). Folding the main chain of small proteins with the genetic algorithm. J. Mol. Bio., **236**, 844-861.

114. Wallqvist A., Jernigan R.L. & Covell D.G. (1995). A preference-based free-energy parametrization of enzyme-inhibitor binding. Applications to HIV-1 protease inhibitor design. Protein Sci., 4, 1881-1903.

115. Verkhivker G, Appelt K, Freer ST, Villafranca JE (1995). Empirical free energy calculations of ligand-protein crystallographic complexes. I. Knowledge-based ligand-protein interaction potentials applied to the prediction of human immunodeficiency virus 1 protease binding affinity. Protein Eng., 7, 677-691.

116. Frisch, M.J., Trucks, G.W., Schlegel, H.B., Gill, P.M.W., Johnson, B.G., Robb, M.A., Cheeseman, J.R., Keith, T.A., Petersson, G.A., Montgomery, J.A., Raghavachari, K., Al-Laham, M.A., Zakrzewski, V.G., Ortiz, J.V., Foresman, J.B., Cioslowski, J., Stefanov, B.B., Nanayakkara, A., Challacombe, M., Peng, C.Y., Ayala, P.Y., Chen, W., Wong, M.W., Andres, J.L., Replogle, E.S., Gomperts, R., Martin, R.L., Fox, D.J., Binkley, J.S., Defrees, D.J., Baker, J., Stewart, J.P., Head-Gordon, M., Gonzalez, C. & Pople, J.A..Gaussian 94 (Revision A.1). (1995). M., Gonzalez, C. & Pople, J.A..Gaussian 94 (Revision A.1). Gaussian, Inc., Pittsburgh PA.

Table of Figures